GLOBAL OPTIMIZING FLOWS FOR ACTIVE CONTOURS

A Thesis
Presented to
The Academic Faculty

by

Ganesh Sundaramoorthi

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical and Computer Engineering

Georgia Institute of Technology
August 2007

GLOBAL OPTIMIZING FLOWS FOR ACTIVE CONTOURS

Approved by:

Professor Anthony Yezzi, Advisor
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Steven McLaughlin
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Allen Tannenbaum
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor John McCuan
School of Mathematics
Georgia Institute of Technology

Professor Patricio Vela
School of Electrical and Computer Engineering
Georgia Institute of Technology

Date Approved: 03 July 2007

To Lord Sri Krishna; may you always engage me in your transcendental service.

ACKNOWLEDGEMENTS

Good research seems to culminate from the combined work of many great persons. To develop into a learned individual, one needs the mentoring, support, and collaboration of great persons. I have a number of great people to thank for their unique contributions to this thesis, and on my personal and professional development. First, I thank my advisor Prof. Anthony Yezzi for mentoring me in every possible aspect of my research. His creativity and patience is a wonder to observe. I thank Prof. Andrea Mennucci, a great mathematician, for his collaborations in which I have learned a lot of interesting mathematics. I thank Professors Allen Tannenbaum, Magnus Egerstedt, Patricio Vela, Steven McLaughlin, Nikil Jayant, and John McCuan for serving on my proposal and/or my dissertation defense committees.

During my stay at Georgia Tech, I had the opportunity to interact with a number of colleagues (friends) with unique personalities. I enjoyed the friendship of Chris Alvino and all the math we discussed. I thank Jeremy Jackson for all the discussions we had about research or otherwise, and of course, for his quick response to every possible Linux question I had! It was a pleasure to work with my other colleagues (friends) past and present: Kelvin Rocha, Hua "Carol" Li, Greg Slabaugh, Gozde Unal, Tejas Mehta, Deryck Yeung, Guillermo Gallego, Siddharth Manay, Miguel Lopez, Gerardo Valdes, Henrik Axelsson, and Abubaker Muhammad. Also, I thank all my other friends in Atlanta and other parts of the world.

I would like to thank my family for providing me the support and the conditions that allowed me to concentrate on studies.

TABLE OF CONTENTS

DEDICATION iii

ACKNOWLEDGEMENTS iv

LIST OF FIGURES ix

SUMMARY xiii

I INTRODUCTION AND BACKGROUND 1

1.1 Curve Evolutions, Shape Optimization, Active Contours 3

1.1.1 Curve Evolution Theory 3

1.1.2 Shape Optimization for Image Segmentation 5

1.1.3 Level Set Evolutions and Existence/Uniqueness of Flows . . 8

1.2 "Shape Priors" to Statistical Shape Analysis 9

1.2.1 Prior Shape Segmentation 9

1.2.2 Statistical Analysis Using Riemannian Manifolds 12

1.3 Contributions of this Thesis 15

II GLOBAL REGULARIZING FLOWS WITH TOPOLOGY PRESERVATION FOR ACTIVE CONTOURS AND POLYGONS 17

2.1 Introduction 18

2.2 Variational Approach to Topology Preservation 22

2.2.1 Polygon Case 25

2.2.2 Differentiable Contour Case 31

2.3 Numerical Implementation 41

2.3.1 Polygons 41

2.3.2 Level Set Method 42

2.3.3 Alternative Implementation using FFT 46

2.4 Simulations 49

2.4.1 Geometric Properties 49

2.4.2 Image Segmentations 51

2.5 Conclusion 55

III SOBOLEV ACTIVE CONTOURS 57

3.1 Introduction 57

3.2 General Theory 61

3.2.1 Structure on the Space of Curves 61

3.2.2 Inner Products on T_cM 63

3.2.3 Gradient of Functionals on M 64

3.2.4 Relation Between H^n and $\tilde{H}^n$ 66

3.3 H^1 and $\tilde{H}^1$ Gradients 68

3.3.1 Solving the ODEs 69

3.3.2 Properties of the Kernels 70

3.3.3 Properties of Sobolev Gradients 71

3.3.4 Advantages of $\tilde{H}^1$ over H^1 73

3.4 Some Sobolev Gradient Flows 74

3.4.1 Length and Weighted Length 75

3.4.2 Area and Weighted Area 76

3.4.3 Elastic Energy 77

3.4.4 Comparison of H^0 and H^1, $\tilde{H}^1$ 79

3.5 Numerical Implementation 80

3.5.1 Multiple Curves Evolution 80

3.5.2 Level Set Method 83

3.5.3 Level Set Method Without Polygon Extraction 85

3.6 Experiments 87

3.6.1 Merging and Splitting 88

3.6.2 Noisy Square Segmentation 89

3.6.3 Segmentation of Real Images 90

3.7 Conclusion 93

IV COARSE-TO-FINE SEGMENTATION AND TRACKING USING SOBOLEV ACTIVE CONTOURS 96
4.1 Introduction 97
4.1.1 Related Works 100
4.2 Some Motivation for Sobolev Metrics 104
4.2.1 Consistent Theory of Shape 104
4.2.2 Fewer Numerical Local Minimizers 106
4.3 Scale-Space Type Analysis of Sobolev Active Contours 107
4.3.1 Sobolev Norms in Frequency Domain 108
4.3.2 Sobolev Gradients in Frequency Domain 109
4.3.3 Coarse-To-Fine Motion of Sobolev Contours 110
4.3.4 Analytic Examples of Coarse-To-Fine Motion of Sobolev Flows 112
4.4 Benefits of Sobolev Contours for Tracking 120
4.5 Experiments 123
4.6 Conclusion 130
V FURTHER APPLICATIONS OF SOBOLEV ACTIVE CONTOURS . . 132
5.1 Introduction 132
5.2 Some Useful Energies Precluded by H^0 134
5.3 Geometric Priors for Active Contours 137
5.3.1 Length and Smoothness Priors 137
5.3.2 Centroid and Isoperimetric Priors 138
5.4 New Edge-Based Active Contour Models 140
5.4.1 Non-Shrinking Edge-Based Model 140
5.4.2 Increasing Weighted Length 141
5.5 Experiments 142
5.5.1 Regularity of Sobolev Active Contour 142
5.5.2 Tracking with Centroid/ Isoperimetric Prior 142
5.5.3 Edge Detection with Non-Shrinking Model 144
5.5.4 Edge Detection by Increasing Weighted Length 145

5.6 Conclusion 146
VI CONCLUSION 147
APPENDIX A CALCULATIONS FOR CHAPTER 2 150
REFERENCES 162
VITA 171

LIST OF FIGURES

1 Illustration of three different cases of polygon topology change or self intersection are shown in (a), (b), (c). Definition of quantities used in justifying Proposition 2.2.1 is given in (d). 29

2 A curve moving toward topology change with quantities labeled as in the justification of Conjecture 2.2.1 is shown in (a). A topology change cannot occur, that is $C(p_2(t), t) \nrightarrow C(p_1(t), t)$, if locally, the curve and directions are indicated in (b) for all time close to t_i. (c) shows a diagram for quantities defined in Step 4 and 5 of the proof. 37

3 Snapshots of an evolution of a polygonal spiral (top) under (25), and evolution of a spiral (bottom) under (45). The polygon is evolving solely under the proposed topology preserving flow. For visibility, the spatial scale is shrunk as the evolution progresses. 50

4 Illustration of a difference between curvature flow (top) and proposed flow (45) (bottom). 51

5 Snapshots of an evolution of active polygon using Chan-Vese flow with the proposed topology preserving flow. 52

6 Segmentation of bones CT image with no topology preservation (left), with topology preservation of Han et al. alone (middle), and with proposed topology preservation (right). 53

7 Gap between bones with contours overlayed. Results of no topology preservation, topology preservation of Han alone, and proposed topology preservation. 53

8 Geometrical pathology typical of forced topology preservation constraint of Han et al. alone (top), and the correct segmentation by the proposed geometric regularization flow. 54

9 The effect of changing the weight on the topology preserving term. (a) Initial contour, (b) and (f) result of Han *et al.*, (c) and (g) 33%, (d) and (h) 20%, (e) and (i) 7% weights on proposed topology preserving term. (f), (g), (h). (i) are enlargements of the middle area between the circles. 55

10 Plots of K_λ (left) and $\tilde{K}_\lambda$ (right) for various λ with $L = 1$. The plots show the kernels over one period. 71

11 Some illustrations of topological changes. 81

12 Segmentation of various shapes using a Sobolev active contour for a region-based energy (89). This illustrates the ability of Sobolev active contours to cope with topological changes: merging and splitting. . . 88

13 Segmentation of square binary image with salt and pepper noise of various densities. The experiment shows the results with H^0 (of various degrees of regularization, α) and the Sobolev active contour. 91

14 Segmentation of a cardiac image using the edge-based energy in (88) (two different initializations). 92

15 Segmentation of an ultrasound image using the region-based energy (89). 93

16 Segmentation of a vessel image using the region-based energy (89) with both the H^0 active contour and the Sobolev active contour. The Sobolev active contour is able to avoid distracting fine features of the image and therefore does not leak into an irrelevant region of the image. 94

17 Top row: H^0, middle: translation favored H^0, and bottom: Sobolev flows for (90). The initialization is a "local" minimum wrt to H^0 of the energy (90) (this is $\hat{c}_0$ as noted in the text). Noise: salt and pepper (density = 0.6). 108

18 Increasingly higher frequency perturbations applied to a circle (left to right, $l = 0, 2, 5, 10$). 111

19 Standard (H^0) active contour (top) alters fine structure of the curve immediately; Sobolev (H^1) active contour (bottom) moves from coarse to finer scale motions. Both use same energy. 112

20 Left: Plot of $\tilde{g}_n(\cdot, t)$ with $n = 2$ for various t, which shows coarse-to-fine behavior. Right: Plot of frequency component changing most rapidly versus time, which shows that the evolution transitions through *all* possible frequencies. 117

21 Minimizing E_m using a "spiked circle" (which is shown on the right of each row) for c_0, and the initialized curve is a circle: $c(u, 0) = \epsilon(\cos 2\pi u, \sin 2\pi u), \epsilon = 0.001$ for $u \in S^1$ (left of each row, enlarged for visibility). Top: Snapshots of the H^0 evolution. Bottom: Snapshots of the $\tilde{H}^2$ evolution. The right of each row shows that both flows converge to the desired "spiked" circle, but taking quite different paths. 118

22 Simple tracking using geodesic active contours: Standard (H^0) active contour (left column) deforms the initialized contour greatly and is stuck in local minima, and Sobolev active contour (right column) moves in a global manner only slightly changing shape. In each frame, the initial curve (given by the contour detected in the previous frame) is blue, the intermediate curve is green, and the final detected curve is red. 122

23 Graphs showing number of iterations to converge and total time for convergence versus set symmetric difference (SSD) of initial region and desired object in percent (scaled by desired object area). 123

24 Tracking of a person in a noisy image sequence with a region-based (Chan-Vese) energy with H^0 (top row), with H^0 translation favored (middle row), and with $\tilde{H}^1$ (bottom row) active contours. 125

25 Plots of error for tracking a square that is translating and slightly changing its area with various degrees of noise (Gaussian mean 0, standard deviation σ). Left: Using geodesic active contours, right: using the Chan-Vese model. Note the difference in the scales of each plot; in particular, the plots show that the results of using Sobolev active contours is vastly better than the corresponding H^0 active contour. . 126

26 Tracking of a sea-creature at the sea bottom using an energy which incorporates the mean intensity and variance information inside and outside the contour. 128

27 Tracking of a car under an occlusion using the Mumford-Shah energy with H^0 (top) and $\tilde{H}^1$ active contours. 129

28 Tracking a car under an occlusion using estimation with Mumford-Shah energy functional for the detection. H^0 (top) and $\tilde{H}^1$ (bottom) active contours. 130

29 H^0 regularization (top two rows). Left to right: $\alpha = 1000$, $\alpha = 1000$ followed by curvature smoothing to remove the noise (least number of iterations to remove noise), $\alpha = 10000, 50000, 90000$. The image-based term is Chan-Vese. Sobolev elastic regularization (bottom two rows). Left to right: $\alpha = 0, 0.1, 5, 10, 25$. The second and fourth row show the same result as the row above them, but the image is removed for visibility. 143

30 Tracking a man through an occlusion. Bottom row shows the results of using a prediction (filtering) on the centroid and the isoperimetric ratio, and then penalizing deviations of the contour away from predicted parameters by (115) ($\alpha = 50000$, $\beta = 100$). The top row gives the result with no such penalty. Both use Sobolev active contours. 144

31 Segmentation of cyst image with three different initializations (first image in each row). Converged results for the (117) and H^0 active contour (second image), (117) with the Sobolev active contour (third image), and the energy (118) (last image). 145

32 Left to right: initial contour, minimizing (121) $\alpha = 0.2, 0.25, 0.4$ using H^0, and increasing weighted (119) $\alpha = 0.1$ using Sobolev (all images show converged contour). The contour expands to enclose the entire image (fifth image). 146

33 Configuration of segments and vertices assumed in the derivations of non-adjacent segment quantities. 151

34 Names of various distances, segment lengths and angles for non-adjacent segments. 151

SUMMARY

This thesis makes significant contributions to the object detection problem in computer vision. The object detection problem is, given a digital image of a scene, to detect the relevant object in the image. One technique for performing object detection, called "active contours," optimizes a constructed energy that is defined on contours (closed curves) and is tailored to image features. An optimization method can be used to perform the optimization of the energy, and thereby deform an initially placed contour to the relevant object. The typical optimization technique used in almost every active contour paper is evolving the contour by the energy's *gradient descent flow*, i.e., the steepest descent flow, in order to drive the initial contour to (hopefully) the minimum curve. The problem with this technique is that often times the contour becomes stuck in a sub-optimal and undesirable local minimum of the energy. This problem can be partially attributed to the fact that the gradient flows of these energies make use of only *local* image and contour information. By local, we mean that in order to evolve a point on the contour, only information local to that point is used. Therefore, in this thesis, we introduce a new class of flows that are *global* in that the evolution of a point on the contour depends on global information from the entire curve. These flows help avoid a number of problems with traditional flows including helping in avoiding undesirable local minima.

In Chapter 2, we demonstrate how one obtains a global flow through the traditional gradient descent technique by constructing a new class of energies that are based on particle interactions for points on the contour. These result in gradient flows that are *global*, and avoid undesirable configurations of the contour that are

local minima of previous energies. We show a particular application of these energies by introducing *global regularity* and topology preserving properties into active contour evolutions. More specifically, we construct a term that can be added to *any* image-based evolution to obtain an evolution that has topology preserving and regularity properties. Such a term has not been constructed before for active contours. These new constructed evolutions provide applications where prior information on the topology and regularity of the object to be detected is known. For example, one application is brain cortex segmentation where the topology of the cortex is known to be topologically equivalent to a sphere, and that topology must be maintained throughout the evolution.

In Chapters 3, 4, and 5 we obtain *global* flows as optimizing flows for all previous active contour energies by considering a novel optimizing technique, which can be applied to any active contour energy. We note that "gradient flows" not only depend on the energy they minimize, but also on the *metric* that is chosen on the space of shapes (contours), i.e., the "cost" to deform contours. This fact has been ignored in the active contour literature, where the agreed upon choice of "metric" has always (knowingly or unknowingly) been the standard geometric L^2-type metric. On the other hand, in a parallel field called shape analysis where the objective is to be able to perform statistical operations on shapes (e.g. contours) there have been numerous metrics considered, and there is no agreed upon choice of metric. A recent result in shape analysis indicates that the standard geometric L^2-type metric considered for defining gradient flows is pathological. As we show, there are also many undesirable properties of the gradient flows for active contour energies that the L^2 metric induces. Therefore, we consider a new class of Sobolev-type geometric metrics to define gradient flows for active contours. These new metrics do not have the pathological properties of the standard L^2-type metric. In addition to the many useful theoretical properties of these new Sobolev-type gradient flows, we show many applications to

segmentation and visual tracking, and the substantial improvements over L^2 gradient flows, in particular avoiding local minima.

CHAPTER I

INTRODUCTION AND BACKGROUND

This thesis concerns the problem of object detection (e.g. segmentation) in computer vision. In other words, given a digital image of a scene, the object detection problem is the problem of designing a computer algorithm to detect the "relevant" object(s) in the image. The approach taken in a large body of literature in computer vision (and that we take in this thesis) is to construct a clever image dependent energy or cost functional defined on the space of all possible objects, that is a space of *shapes*, and find the optimum object or shape with respect to the defined energy. The optimum shape with respect to the constructed energy is the relevant object that is to be detected. The previous problem is known as the *shape optimization* problem, and poses many important questions such as

1. how does one mathematically represent shapes?

2. how does one construct energies on shapes?

3. how does one find the optimum shape of an energy?

In many cases the information in the image itself is not enough to detect the object (e.g. in the case of occlusions of the desired object or missing information from the image). Therefore, many works take the approach of incorporating prior known information on the shape of the object to be detected into the construction of the energy. Such prior information may be available from a database of likely shapes. When such a database is available, it is of interest to quantify differences in shapes so that, for example, it is possible to compare a shape obtained from image data alone to shapes in the database in order to know whether the obtained shape is "likely."

More generally, it may be desired to perform statistical operations on shapes in the database in order to, for example, obtain efficient representations of the database for real-time algorithms. These ideas raise many fundamental questions:

1. how does one mathematically represent shapes?

2. how does one define and construct a metric or similarity score between any two shapes?

3. how does one perform statistical operations on shapes?

This thesis relates to all of the questions above, but we are mainly interested in the third question of the first paragraph. Indeed, we simply follow a common approach in the literature in order to answer the first question. This thesis develops a novel technique to optimize energies defined on shapes and shows how this technique is naturally connected to the question of constructing metrics defined on shapes. As we shall see, the approach we take in this thesis is in some sense a generalization of calculus and statistics in $\mathbb{R}^d$. In $\mathbb{R}^d$, taking a point $x \in \mathbb{R}^d$ to represent a "shape," it is quite easy to answer the previous questions. For example, if we have an energy or cost functional, $E : \mathbb{R}^d \to \mathbb{R}^+$, then an approach for minimizing E is to derive the steepest descent or gradient descent flow:

$$\begin{cases} \dot{x}(t) = -\nabla E(x(t)) = -\left(\dfrac{\partial E}{\partial x_1}(x(t)), \ldots, \dfrac{\partial E}{\partial x_d}(x(t))\right) \\ x(0) = x^0 \end{cases}$$

where $x^0 \in \mathbb{R}^d$ is an initial guess of the minimum of E, $x(t) = (x_1(t), \ldots, x_d(t))$, and $\dot{x}$ is the time derivative of x. If the flow exists, then it will converge to a local minimum of E (provided one exists). A metric on $\mathbb{R}^d$ can simply be taken to be the Euclidean metric. We can easily define statistics in $\mathbb{R}^d$: indeed the average of a set of points $y_1, \ldots, y_k \in \mathbb{R}^d$, where k is a positive integer, is

$$\hat{y} = \frac{1}{k}\sum_{i=1}^{k} y_i.$$

Hidden in all of the above simple calculus and statistics computations is the notion of "metric," and we shall see in this thesis how the metric choice on more complicated spaces of "shapes" is crucial for the object detection problem.

In the next sections, we give a review of the relevant literature, which will in part help motivate the contributions of this thesis.

1.1 Curve Evolutions, Shape Optimization, Active Contours

Let $\Omega \subset \mathbb{R}^2$ denote the domain of an image $I : \Omega \to \mathbb{R}^+$. We are interested in image segmentation: partitioning Ω into subsets A_i such that $A_i \cap A_j = \emptyset$ when $i \neq j$, $\cup_i A_i = \Omega$, and each A_i should represent a "distinctive" region of the image I. In this thesis, we consider the simplest case of two regions: A_1, representing an object of interest in the image (e.g. a car or person), and A_2, the background. One approach, which we take in this thesis, is done by designing an algorithm to *deform* an initially placed *shape* in Ω to the relevant object in I. In order to deform a "shape", one first needs a computational definition of shape; unfortunately, there is no agreeable definition (see Section 1.2). However, in the computer vision literature a somewhat natural and intuitive definition has arisen, although not explicitly stated, from the works of Kass et al. [51] and Mumford and Shah [71, 72]: a shape is simply a region $A \subset \Omega$ represented by its bounding contour. This definition is made explicit in the work of Kimia et al. [55] in which the authors formulate a theory of shape based on the theory of deformations from the mathematical literature. Indeed, the authors represent shapes as *images* of C^2 embedded mappings $S^1 \to \mathbb{R}^2$, i.e., curves that are diffeomorphic to the circle S^1. We will build from this approach in this thesis.

1.1.1 Curve Evolution Theory

The origin of the active contour field, which is generally a subset of shape optimization where "shape" means "contour," can be traced to the mathematical literature on curve evolutions, in particular, the curve shortening flow. The curve shortening

flow is the flow obtained by deforming a smooth curve at a velocity proportional to its curvature vector. The term "curve shortening" originated in the mathematical literature "because the flow lines in the space of closed curves are tangent to *the gradient* for the length functional [41]." Mathematically, the curve shortening flow is defined by the following non-linear partial differential equation (PDE):

$$\begin{cases} \dfrac{\partial C}{\partial t}(u,t) = \dfrac{\partial^2 C}{\partial s^2}(u,t) = \kappa(u,t)\mathcal{N}(u,t) \\ C(u,0) = c_0(u) \end{cases} \tag{1}$$

where $u \in S^1$, $c_0 \in C^2(S^1, \mathbb{R}^2)$ is embedded, $C : S^1 \times \mathbb{R}^+ \to \mathbb{R}^2$ is a family of curves, κ is the curvature of C, $\mathcal{N}$ is the inward unit normal, and $s : S^1 \times \mathbb{R}^+$ is the arc-length parameter, which is defined as

$$s(u,t) = \int_0^u |\partial_u C(u,t)| \, \mathrm{d}u. \tag{2}$$

In a series of papers [35, 36, 37, 40], Gage and Hamilton, and Grayson prove remarkable results about (1): the curve $C(\cdot, t)$ remains an embedded curve, and the initial curve c_0 is shrunk to a "circular point" in finite time in the sense that the isoperimetric ratio of $C(\cdot, t)$ converges to that of a circle. Thus, this equation generalizes Gaussian smoothing to geometric curves in the sense that Gaussian smoothing arises from a linear heat equation on the plane $\mathbb{R}^2$, which resembles (1). It should be noted that the curve shortening flow is invariant to the Euclidean group in the sense that if C is the solution to (1) then the solution of (1) in which c_0 is replaced with $g \circ c_0$ where g is a Euclidean action is $g \circ C$. In computer vision, it is of interest to have a similar smoothing process that is affine invariant; indeed, the flow

$$\frac{\partial C}{\partial t}(u,t) = \kappa^{1/3}(u,t)\mathcal{N}(u,t)$$

considered by Sapiro and Tannenbaum [91, 92] is affine invariant and has similar smoothness and embeddedness properties to the curve shortening flow.

In computer vision, Kimia et al. [55] have introduced the flow

$$C_t = (\beta_0 + \beta_1 \kappa)\mathcal{N}, \tag{3}$$

where $\beta_1 > 0$, for the purpose of shape analysis. The flow was numerically studied earlier by Osher and Sethian [78] and Sethian [94]. This PDE is known as a *reaction-diffusion* equation. The first term is an inflationary term moving the curve outward and produces singularities of the curve, and the second term is a diffusion that smooths the curve as noted earlier. Caselles et al. [14, 15] and Malladi et al. [63] have generalized (3) to perform boundary detection for the image segmentation problem. They consider the curve evolution

$$C_t = \phi(\kappa - \nu)\mathcal{N}, \tag{4}$$

where $\nu \in \mathbb{R}$ is a constant, which controls the inward or outward (depending on the sign of ν) movement of the curve, and $\phi : \Omega \to \mathbb{R}^+$ is an "image edge detector," which is low near object boundaries in the image; typically, one chooses

$$\phi(x) = \frac{1}{1 + |\nabla I(x)|^2}, \text{ where } \nabla I(x) = (\partial_{x_1} I(x), \partial_{x_2} I(x)). \tag{5}$$

The curvature term ensures curve regularity. In the case $\nu > 0$, the curve according to (4) moves outward smoothly, and the speed of the curve is controlled by ϕ. When the curve becomes close to edges, ϕ is small and the curve slows down and comes to a stop.

1.1.2 Shape Optimization for Image Segmentation

An energy based technique for evolving curves to detect an object in an image was proposed in the seminal work of Kass et al. [51]. Minimizing the energy or cost functional, $E : C^2(S^1, \Omega) \to \mathbb{R}^+$, defined on closed curves given by

$$E(c) = \int_{S^1} \left(\phi(c(u)) + w_1(u)\|c'(u)\|^2 + w_2(u)\|c''(u)\|^2\right) \mathrm{d}u \tag{6}$$

was proposed to evolve an initial curve to the target object boundary. Note that $w_1, w_2 : S^1 \to \mathbb{R}^+$ are predetermined weights, and ϕ is the edge detector mentioned earlier (5). The idea is that the energy is minimized by curves that align along the

boundary of the object (since ϕ is low in these places) and are smooth because the second two terms of (6) are small for smooth curves. To minimize (6), one calculates its Euler-Lagrange equation and constructs *the gradient flow* to determine a curve evolution that will lock into a local minimum of (6). We outline such a procedure in the next paragraph.

One problem with the model of Kass et al. [51] is that it is dependent on the parameterization of c. This is clearly undesirable for image segmentation applications. The work of Kichenassamy et al. [53] and Caselles et al. [16] (see also the work of Shah [95]) united the curve evolution approaches [14, 15, 63] and the energy minimization approach to object detection of [51] by proposing to minimize the following weighted length energy, which is *geometric* and is similar to (6):

$$E(c) = \int_c \phi(c(s))\,\mathrm{d}s \tag{7}$$

where s is defined in (2) and $\mathrm{d}s$ is the arc-length measure. Note that (7) is invariant to parameterization of c since only the arc-length parametrization is considered. One can compute the Euler-Lagrange equation by computing the variation of E to find

$$\frac{\mathrm{d}}{\mathrm{d}t}\, E(c+th)|_{t=0} = \int_c h \cdot (\nabla\phi \cdot \mathcal{N} - \phi\kappa)\mathcal{N}\,\mathrm{d}s \tag{8}$$

where $h : S^1 \to \mathbb{R}^2$ is an infinitesimal deformation of c, and $(c+th)(u) := c(u)+th(u)$. The variation above is zero for all smooth h when

$$(\nabla\phi \cdot \mathcal{N} - \phi\kappa)\mathcal{N} = 0,$$

which is the Euler-Lagrange equation for (7). To minimize (7), one can consider evolving an initial curve according to the curve evolution

$$C_t = -(\nabla\phi \cdot \mathcal{N})\mathcal{N} + \phi\kappa\mathcal{N}, \tag{9}$$

which according to (8) minimizes E, and is referred to as *the gradient descent flow*. The first term attracts the the curve to the object boundary, while the second term

shrinks the curve and slows near the object boundary. Note that (9) generalizes the curve shortening flow (1), which is the gradient flow for length, to a more general Riemannian manifold (instead of the isotropic plane) whose metric is determined by ϕ.

The common feature of the above curve evolutions for image segmentation is that they use information from image edges, determined by ϕ, in order to move the curve. Other edge-based models include [24, 111, 98]. For many images, this edge information is not sufficient to drive the curve to segment the image correctly. Many works (e.g. [71, 72, 88, 118, 114, 17, 80, 81]) have incorporated region-based information (e.g. statistical properties of image regions) to derive curve evolutions. One of the first energies using region-based information is the energy considered by Mumford and Shah [71, 72] (related to the work of Geman and Geman [38], and Blake and Zisserman [7]):

$$E(c, f) = \int_{\Omega} (I - f)^2 \, \mathrm{d}A + \alpha \int_{\Omega \setminus c(S^1)} |\nabla f| \, \mathrm{d}A + \beta \int_c \mathrm{d}s \tag{10}$$

where $\alpha, \beta > 0$, I is the image to be segmented, $f : \Omega \to \mathbb{R}^+$ is a piecewise smooth function that may be discontinuous along c, and $\mathrm{d}A$ is the area measure in Ω. The energy is minimized in both c and f. It is optimized when the curve c separates distinct regions of the image, and f smoothly approximates the image I inside and outside the curve c. Mumford and Shah [72] prove existence of a minimizer of (10) using geometric measure theory when f is restricted to be piecewise constant. Other region-based curve energies (e.g. [88, 118, 114, 17, 80, 81]) are based on minimizing the weighted area energy defined on embedded curves:

$$E(c) = \int_{c_{in}} g \, \mathrm{d}A, \tag{11}$$

where c_{in} is the region enclosed by c and $g : \Omega \to \mathbb{R}$. *The gradient flow* of (11) (see for example [118]) is

$$C_t = g\mathcal{N}. \tag{12}$$

1.1.3 Level Set Evolutions and Existence/Uniqueness of Flows

We now discuss the numerical implementation of the curve evolutions discussed in the previous sections. In the work of Kass et al. [51], the curve evolution is implemented using a Lagrangian marker particle method, i.e., the curve is represented by sample points, and the flow is discretized. While the method is quite simple, there is a basic problem: the "particles" become close during the evolution, which leads to many numerical problems. Therefore, the approach taken by most of the authors of the works discussed in the previous sections is using *level set methods*, introduced by Osher and Sethian [78]. In this approach a function $\Psi : \Omega \times \mathbb{R}^+ \to \mathbb{R}$ called the level set function is evolved rather than the curve C. The zero-level set of Ψ is C, i.e., $\{x \in \Omega : \Psi(x,t) = 0\} = C(S^1, t)$. Note that the evolution of Ψ is

$$\Psi_t(C(u,t)) = -\nabla\Psi(C(u,t)) \cdot C_t(u,t) \tag{13}$$

for points on the curve C. Note that the previous equation only specifies the evolution of Ψ along the curve C, and therefore the evolution must be specified for other points $x \in \Omega$. For many of the non-image based evolutions (e.g. (1),(3)) there is no problem, and the evolutions can be naturally extended to all of Ω. For example, if we consider the curve evolution $C_t = \mathcal{N}$, then we can write $\mathcal{N} = -\nabla\Psi/|\nabla\Psi|$, which is defined on Ω not just C, when Ψ is restricted to be negative inside C and positive outside. Then (13) becomes $\Psi_t = -|\nabla\Psi|$, which is defined on Ω. For image-based evolutions, one usually has to construct *extensions* in order to define the level set evolution. The benefits of level set methods are numerous: including numerical stability, not having to deal with reparameterizations, and splitting and merging of curves are handled automatically. However, the method may only be used on certain curve evolutions, namely, evolutions that have a *maximum principle*.

Level set methods are not only advantageous for numerical implementation, but also for theoretical reasons. Indeed the corresponding level set evolution equations are

generally much easier to analyze than the corresponding geometric curve evolutions mainly because level set functions form a linear space. The works [35, 36, 37, 40] have shown the difficulty of proving existence/uniqueness of a simple geometric curve evolution. Therefore, many works (e.g. [53, 16]) have taken the approach of converting the curve evolution into a level set equation, and then proving existence/uniqueness of this PDE in some sense. There is no general theory for these level set evolutions, which are almost always non-linear PDE, and each level set equation typically must be analyzed using different techniques. Viscosity theory [25] and the theory of hyperbolic conservation laws provides a convenient setting for analyzing such non-linear PDE. This theory provides a *weak* formulation of the level set PDE, and accounts for non-smooth solutions. We shall not go into details of this theory. Another advantage of the level set evolutions over corresponding curve evolutions is that they make sense theoretically when topology changes occur; such changes would result in cusps or kinks in the curve, and thus the curve evolution equation would not make sense.

1.2 "Shape Priors" to Statistical Shape Analysis

1.2.1 Prior Shape Segmentation

Recent works that use active contours for segmentation are not only based on information from the image to be segmented (edge-based or region-based), but also on *a priori* information known about the *shape* of the desired object to be segmented. The need for this type of prior term arose from several factors including the fact that some images contain limited information, the energies functions considered previously could not incorporate complex information, the energies had too many extraneous local minima, and the *gradient flows* to minimize these energies allowed for arbitrary deformations that gave rise to unlikely shapes.

The work of Leventon et al. [60] showed how to incorporate the information from a database of shapes that represent different variations of the object to be segmented

into the active contour paradigm. In this work, it is assumed that a database of shapes represented by their *distance* functions is available. A distance function to a curve c, $d_c : \Omega \to \mathbb{R}$, is simply defined by

$$d_c(x) = \inf_{y \in c(S^1)} |x - y|$$

where $c(S^1)$ is the image of c. Usually one considers the signed distance function, which is identical to d_c except that the signed version is negative when x is in the region enclosed by c. Therefore, signed distance functions are special cases of level set functions. The idea in [60] is to perform a principal component analysis (PCA) on the signed distance functions of the database shapes, which can be discretely represented as column vectors. Although the set of signed distance functions is not a vector space, it is shown by experiments that the zero level sets of linear combinations of the components obtained from the PCA "look like" reasonable shapes. An alignment of the shapes so that the shapes have nearly identical pose is done by an *ad-hoc* method before the PCA. The procedure for the authors' segmentation using active contours is to evolve the initial curve for a few iterations by the geodesic active contour model [53, 16], and then move the contour according to the prior shape information. This procedure is continued until convergence. In order to move the contour toward to a *probable* shape, a shape, which is determined by the PCA, is obtained by maximizing the probability of it given the current contour and image intensity information. The contour is then moved toward this maximum probability shape by simply adding to the contour a scaled version of the shape.

Subsequently, Tsai et al.[107] built on the work of [60]. Instead of the *ad-hoc* procedure for aligning training data, Tsai et al.propose an alignment procedure by minimizing a simple energy defined on *pose* parameters for each training image. The energy is simply a sum of the normalized L^2 discrepancy between mutual pairs of pose transformed images. They then perform a PCA on signed distance functions of the training data to obtain the components $\{\Phi_0, \ldots, \Phi_{k-1}\}$. A shape is modeled as

the zero level set of

$$\Phi(w, T) = \Phi_0 \circ T_0 + \sum_{i=1}^{k-1} w_i \Phi_i \circ T_i$$

where $w = \{w_i\}_1^{k-1} \subset \mathbb{R}$ are weights, and $T = \{T_i\}_0^{k-1}$ are pose transformations represented by a finite number of parameters. Unlike the usual active contour strategy that attempts to find the best contour for a given energy, the authors of [107] optimize a traditional active contour energy, for example a region-based energy discussed earlier, with respect to the weights, w, and pose parameters, T. Therefore, the shape model is directly incorporated into the image-based segmentation unlike the approach of [60]. The problem with the approach of [107] as with that of [60] is that the set of signed distance functions is not a vector space. Thus, Duci et al.[32] considered a representation of shapes that gives a linear structure. Other works that are related to [107] are [89, 22, 27, 85]. In all these works, the main idea is designing a term of the energy to incorporate prior shapes that is small when the curve is *close*, in some sense, to a pre-specified shape.

A problem with the methods discussed in this section, among others, is the need for training data, which is sometimes unavailable. Therefore, there have been some works that attempt to incorporate prior shape information to the segmentation energy functional by adding a term that is favorable to certain geometries of shapes without the use of training data. For example, the work of Rochery et al. [86] attempts to segment line structures in satellite images. These line structures are thin elongated structures, and a prior term is created to favor these shapes. They introduce a new class of active contours energies, which in particular favor thin elongated structures. Another example of a geometric shape prior driven term is the work of Nain et al. [73], which attempts to segment vessels.

1.2.2 Statistical Analysis Using Riemannian Manifolds

The works on shape priors presented in Section 1.2.1 have led to a fundamental question on how to define a distance between two curves or shapes. Such a distance is useful for many other applications in computer vision as well. This distance allows one to define an average shape of a given set of shapes, and more generally compute statistics of shapes. One example application is the problem of object classification where one may have a database of known shapes and it may be desirable to classify a new shape. Determining whether a new shape is part of an existing class can be performed when statistics on shapes can be computed. Another example is the registration problem where one wants to determine the morph transforming one curve to an another; as we shall see, this problem is related to defining distance between shapes. This area of research called *shape analysis* spans at least a hundred years, although the applications of shape analysis to prior-based segmentation with active contours has only been introduced in the past few years.

Early works on defining shape spaces in the statistics literature are found in [52, 58, 65]. In these works, shapes are modeled as a quotient space of sets of N points in $\mathbb{R}^M$; that is, the shape space is $\mathbb{R}^{MN}/\{SE(M) \times \mathbb{R}^+\}$ where $SE(M)$ represents the special Euclidean group, and $\mathbb{R}^+$ represents the possible scale parameter. Transformations of a shape within $SE(M) \times \mathbb{R}$ are considered to generate the same shape, while those outside correspond to deformations of the shape. This space is endowed with a Riemannian metric to quantify differences of shape. The disadvantage of this approach is that all shapes must have N distinct parts or landmarks, and one cannot compare shapes that do not have distinct landmarks. Grenander [42] gives an approach for defining a shape space without the use of landmarks. In this work, the shape space is represented as an infinite dimensional manifold, and variations of shape are captured by actions of Lie groups. Low dimensional transformations (i.e., $SE(n)$) leave shapes unchanged, while high dimensional transformations deform shapes.

We now highlight some recent works in shape analysis, which build from Grenander's theory [42] of using infinite dimensional Riemannian manifolds to model shape spaces, and which we shall build on in this thesis. Younes [117] builds on the work of Grenander [43, 42] by defining shapes as *elastic curves*. The distance between these curves is not based on a discretization of the curve or landmarks. Curves are defined by their length, and their tangent vector field. The cost of perturbing a curve, which is constructed to be invariant with respect to translations is related to a Sobolev norm. A first order approximation of this cost is used to construct the distance between two curves. The distance construction is based on the distance generated by a Riemannian manifold. A numerical implementation is given for the case when the curves are restricted to polygons. This theory is generalized in the work of Trouvé and Younes [105], in which the authors construct a Riemannian manifold by considering spaces of images (i.e., the shapes) and deformations of the entire plane as infinitesimal shape variations. It is shown that this construction simplifies to both the case of Riemannian manifolds based on landmarks and the elastic curves manifold of [117] in special cases.

Mio et al. [70] represent shapes as curves, and curves are represented as *direction functions*, that is, angle functions of tangent vectors. The shape space is produced from $\theta_0 + L^2(S^1, \mathbb{R})$ where $\theta_0(s) = s$ is the angle function for the circle. They would like to produce a space invariant to rotation, and this is done by restricting to functions $\theta \in \theta_0 + L^2$ such that the average of θ is π. Moreover, for a θ to represent a closed curve, it must satisfy: $\int_0^{2\pi} \exp i\theta \, \mathrm{d}s = 0$. Therefore, they define the map $f : \theta_0 + L^2 \to \mathbb{R}^3$

$$f^1(\theta) = \frac{1}{2\pi} \int_0^{2\pi} \theta(s) \, \mathrm{d}s, \; f^2(\theta) = \int_0^{2\pi} \cos\theta(s) \, \mathrm{d}s, \; f^3(\theta) = \int_0^{2\pi} \sin\theta(s) \, \mathrm{d}s,$$

and the pre-shape sub-manifold is $\mathcal{C}_1 = f^{-1}(\pi, 0, 0)$. Note that the space has a dependence on the initial point, and so the quotient space $\mathcal{S} = \mathcal{C}_1 / S^1$ is considered the shape space. This shape space is a Riemannian manifold, and the authors give an approximate method for computing geodesics on this manifold by approximating

curves through the Fourier basis.

There have been numerous works in the shape analysis literature, and each author has proposed different way of defining distance between shapes. This can be partly attributed to the fact that human perception of shape similarity is so complex, and a single simple mathematical metric cannot possibly be sufficient to describe human perception. Michor and Mumford [68] analyze the simplest possible Riemannian metric on plane curves. This motivation leads the authors to consider the L^2-type geometric inner product, which they referred to as H^0, on the space of deformations of a curve. However, [68] (independently by [112]) show a rather remarkable property about the Riemannian metric on the space of curves induced by the H^0 inner product. Indeed, the authors show that the distance between any two curves induced from H^0 is zero, which makes this "distance" meaningless. Accordingly, Michor and Mumford propose a slightly modified version of the H^0 inner product:

$$\langle h, k \rangle_{H^0_A, c} := \int_c (1 + A\kappa^2(s)) h(s) \cdot k(s) \, \mathrm{d}s, \tag{14}$$

where $h, k \in C^\infty(S^1, \Omega)$ are perturbations of the curve c, $A > 0$, and κ is the curvature of c. This induces a non-trivial Riemannian metric on the space of plane curves.

Yezzi and Mennucci [112, 116] also analyze the H^0 Riemannian structure, motivated by the fact that all works in the geometric active contours literature (as described in Section 1.1) that derive *gradient flows* to minimize energies imply a natural notion of distance between curves or shapes, i.e., the distance induced from the H^0 Riemannian structure. Hence, if one would like a *consistent* theory of both shape optimization (e.g. active contours) and shape analysis, then one should use the H^0 "metric" for shape analysis. As noted by Yezzi and Mennucci, the proposed modification (14) of the H^0 "metric," although yielding a non-trivial metric, has detrimental consequences for defining active contours (e.g. loss of maximum principle for the

length shortening flow). Therefore, they propose a conformal variant of H^0:

$$\langle h, k \rangle_{H^0_{\phi},c} := \phi(L(c)) \int_c h(s) \cdot k(s) \, \mathrm{d}s \tag{15}$$

where $L(c)$ is the length of c, and $\phi : \mathbb{R} \to \mathbb{R}^+$ is a conformal factor. This leads to gradient flows that are geometrically equivalent to the corresponding H^0 gradient flows, and moreover the induced Riemannian metric is non-trivial. Unfortunately, recent computations done by Shah [96] suggest that minimal geodesics between curves do not exist in general.

1.3 Contributions of this Thesis

The main contribution of this thesis is to essentially introduce *global* minimizing flows to solve active contour problems. By global flow, we mean that the curve deformation at a particular point at all instants of time for active contours depends on *global* information from the entire curve, not just information local to the point as in all previous active contour models. There are many reasons for introducing these flows, including to avoid *local minima* of energies caused by *unlikely shape deformations*. This thesis also moves a step in the direction of obtaining a consistent theory of both shape optimization and shape analysis, which has been elusive in the past.

More specifically, the contributions of this thesis are the following:

- In Chapter 2, we start by introducing a novel geometric double integral energy for active contours. This introduces a new class of energies that are different than traditional active contour energies, which are often "linear." The gradient descent flow of this energy depends on global information of the curve. We apply this type of energy to construct a new class of active contours with *global regularizing* and topology preserving properties. This can be thought of as a "shape prior" for active contour segmentation. A topology preserving active contour is of great interest for applications such as brain cortex segmentation where the cortex is always homeomorphic to a sphere.

- Next in Chapter 3, we introduce the fact that gradient flows for active contour energies depend on the *metric* that exists on the space of curves (shapes). This fact has been ignored in the active contour literature throughout the 20 year history of the field. We formalize the theory of active contours and introduce the manifold in which they live. We then introduce Sobolev-type metrics in the space of curves. This change of metric changes all previous *gradient flows* for active contours, which have all been (knowingly or unknowingly) based on the traditional H^0 "metric." This gives a completely new way of optimizing any active contour energy. These new gradient flows are *global* flows that in particular avoid local minima of energies that plague H^0 gradient flows. We illustrate and discuss the many useful properties of these new flows, including applications to both segmentation and visual tracking in Chapters 4 and 5. This work also moves in the direction of a consistent theory of shape optimization and analysis as we will show.

This thesis is a combination of the following papers: [100, 101, 102, 103].

CHAPTER II

GLOBAL REGULARIZING FLOWS WITH TOPOLOGY PRESERVATION FOR ACTIVE CONTOURS AND POLYGONS

In some applications for image segmentation using active contours, the topology of the object(s) to be detected from an image is known *a priori*, despite a complex unknown geometry, and it is important that the active contour or polygon maintain the desired topology. In this chapter, we construct a novel geometric flow that can be added to image-based evolutions of active contours and polygons in order to preserve the topology of the initial contour or polygon. We emphasize that, unlike other methods for topology preservation, the proposed geometric flow continually adjusts the geometry of the original evolution in a gradual and graceful manner so as to prevent a topology change long before the curve or polygon becomes *close* to topology change. The flow also serves as a *global regularity* term for the evolving contour, and has smoothness properties similar to curvature flow. These properties of gradually adjusting the original flow and global regularization prevent geometrical inaccuracies common with simple discrete topology preservation schemes. The proposed topology preserving geometric flow is the gradient flow arising from an energy that is based on electrostatic principles. The evolution of a single point on the contour depends on all other points of the contour, which is different from traditional curve evolutions in the computer vision literature.

2.1 Introduction

The active contour field has seen a number of modeling techniques from early edge-based strategies to more sophisticated region-based techniques. All of these energy-based active contour models have two common features: a data fidelity term, which matches the curve to the image, and a term that incorporates prior information of the object to be detected, most often a regularity penalty. One possible way to numerically implement active contour evolutions is using level set methods, introduced by Osher and Sethian [78]. Advantages of this approach are that it is independent of parametrization of the curve, and that splitting and merging, i.e., topology changes, of the evolving curve or curves are handled without additional effort.

Although topological flexibility is an advantage of level set methods in many applications, in some applications, such topologies changes are not desirable [45, 44]. In these applications, the topology of the object to be segmented is known *a priori*, and must be enforced during the evolution. The work of Han et al. [45, 44] examines the segmentation of the cortex of the brain. It is known that the cortex is homeomorphic to a two dimensional sphere, and so this topology must be enforced during the evolution. Han et al. give examples where using a level set method to segment the cortex without enforcing the topology of a sphere gives a wrong segmentation. The method for topology preservation in [45, 44] assumes that the active contour evolution is implemented using level sets. Therefore, it is assumed that the curve is evolving in a finite subset of $\mathbb{R}^2$, i.e., the domain of the level set function. The idea of Han et al.'s algorithm is to detect the grid points of the domain of the level set function that will change sign at each iteration of the level set function evolution. These grid points are the only locations where a topology change can occur. If a topology change will occur at one of these grid points as a result of updating the level set function, then the function value at this grid point is not changed. A condition for detecting topology change at a grid point is derived; the condition is based on

the configuration of the level set function in a small neighborhood of the grid point. One of the advantages of this method is computational speed; the method does not accrue much computational cost compared with the scheme without any topology preservation.

Although the method of Han et al. guarantees that the resulting segmentation has the correct topology, there are some undesirable features resulting from the way this method is used in [45, 44]. The most notable disadvantage of Han *et al.*'s method as presented there is that topology preservation is an abrupt, discontinuous motion that is unnatural. Often times, as we shall see in Section 2.4, one obtains an unnatural segmentation, plagued by *geometric inconsistencies*, which is close to self intersection, and is only separated by a minimal spacing that is determined by the grid spacing. The undesirable features resulting from the method of Han et al. stem from the fact that the underlying partial differential equation (PDE) guiding the evolving contour does not, in general, have any "desire" to preserve topology. Thus, the abrupt stopping of the curve in the method of Han et al. does not in any way simulate the actual behavior of the underlying PDE. It would be justified to use the method of Han *et al.*, as a numerical scheme, in conjunction with a PDE that actually preserves topology much like the use of numerical schemes that are designed to simulate certain features of continuous-time PDEs (for example, conservation laws). Thus, in the case that limitations of the discretized PDE prevent topology preservation in the simulation of the PDE, the method of Han et al. can be used to prevent a discrete topology change. This would actually be a better simulation of the PDE, which actually preserves topology, than not using a designed discrete topology preserving step considering the limiting factors of the discretization.

In the proposed work, we construct a PDE that actually preserves topology. We construct a geometric term that can be added to an existing image-based contour evolution. This novel geometric flow preserves the topology of the contour by continually

adjusting the original flow so that the contour is gradually moved away from changing topology well before the contour becomes close to changing topology. We will show that the flow we construct serves to *globally regularize* the evolving contour and may be used as a contour regularizer that, among other properties, ensures smoothness of the contour. Therefore, the usual curvature smoothness term that is used in most active contour works can be replaced with the proposed geometric flow when topology preservation is needed. Geometric inconsistencies in the resulting segmentation that typically arise from using the method of Han et al. on a PDE that does not preserve topology are not likely with our new flow because of the global regularizing properties and gradual topology preservation of our flow. Because we use a geometric flow, our method is not restricted to level set methods, and any active contour implementation can be used to implement our flow. We also construct a global regularizing and topology preserving term for the case of evolving polygons [10, 108].

We *pair* the discrete numerical scheme of Han et al. (or any other discrete topology preservation scheme) with our continuous-time topology preserving PDE because of limiting factors from the discretization of our PDE. The pairing we make is not so different from the pairing of discrete numerical schemes with appropriate PDEs. For example, it is well known that first order evolution PDEs require an appropriate upwinding discretizing scheme. If one were to use an upwinding scheme for second order PDEs, it would lead to numerical artifacts not representative of the PDE. In the same way, using the numerical scheme of Han et al. on an inappropriate PDE leads to numerical artifacts not representative of the actual PDE. Therefore, our method gives new justification for using the numerical scheme of Han et al. in conjunction with a class of PDEs that actually preserve topology. Because implementing our new PDE preserves topology in continuous time, and the underlying PDE prevents topology change before the curve is close to self-intersection, the proposed method is much less likely to suffer from the disadvantages of using the method of Han et al.

on an inappropriate PDE (see Section 2.3.2.2 for more details).

In order to derive a geometric flow for topology preservation in the presence of image-based forces, we observe a necessary property of the flow. Note that a topology change is a self intersection of the contour. In order for a point on the contour to move away from self intersection, it must have some "knowledge" of the location of the other points on the contour. Thus, the evolution of a point must depend globally on other points of the contour, as opposed to a local dependence as a curvature regularizer. The idea of a contour evolution that depends on each point of the contour has been used in recent works, for example, in the work of Rochery et al. [86, 87]. In the work of Rochery et al., an energy is introduced that is a double integral around a contour of some function. This energy is used to enforce a global shape prior that curves are long and thin. The resulting curve evolution depends on an integral around the contour. A similar idea of energies that are double integrals is also used in the work of Kim et al. [54]. In this case, however, the energy is a double integral over the region enclosed by the contour rather than the contour itself. In this chapter, we introduce an energy that depends on a double integral over the contour, and has natural connections to electrostatics.

Applying principles from electrostatics to computer vision has been done in recent work. For example, the authors in [18] use simulated charge inside the region enclosed by an active contour. The contour moves in response to both the electric potential due to the charge and a potential derived from the image to be segmented. The electric field moves the contour outward, and the field from the image potential pushes the contour toward the object to be segmented. This method cannot be used for topology preservation; in fact, the authors in [18] give results that show merging and splitting of contours is possible with their method. A similar use of electrostatics is used in [50], but again, this method cannot be used for topology preservation.

Recent works that incorporate topological priors into image segmentation are [97,

3]. In [97], an energy is constructed to favor distinct connected components of the contour from intersecting; however, there is nothing in the energy that prevents a single connected component from splitting. In [3], a level set scheme is formulated for preserving topology. This method is different than ours in that it is based on the level set representation of the contour whereas our geometric flow can be used with any implementation of the contour evolution. The key difference is that their level set evolution does not provide global regularization of the contour, which is often necessary in applications, and a major contribution of our work.

2.2 Variational Approach to Topology Preservation

We shall first give a mathematical definition of the active polygons and active contours that we consider in this chapter. Let the domain of the image we wish to segment with a polygon or contour be denoted by $\mathcal{I}$, where $\mathcal{I} \subset \mathbb{R}^2$. An active polygon is a polygon with vertexes $v_0, \ldots, v_{n-1}$, where $n \in \mathbb{N}\backslash\{1, 2\}$, in $\mathcal{I}$ that moves at each instant of time according to the set of ordinary differential equations

$$\frac{\mathrm{d}v_k}{\mathrm{d}t}(t) = I_k(t) + \alpha R_k(t), \text{where } k = 0, 1, \ldots, n-1 \tag{16}$$

where $I_k, R_k : \mathbb{R}^+ \to \mathbb{R}^2$; I_k is the force derived from the image we wish to segment, R_k is a regularizing and topology preserving force that is derived only from the geometry of the polygon, and $\alpha > 0$ is a constant. We assume uniform boundedness of I_k, that is,

$$\sup_{t \in \mathbb{R}^+, k=0,1,\ldots,n-1} \|I_k(t)\| < +\infty, \tag{17}$$

where $\|\cdot\|$ is the Euclidean norm of $\cdot$. An active contour C is a twice differentiable curve in $\mathcal{I}$ that moves at each instant of time according to the partial differential equation

$$\frac{\partial C}{\partial t}(p, t) = i(p, t)\mathcal{N}(p, t) + \alpha\mathcal{R}(p, t) \tag{18}$$

where $p \in S^1$ denotes a parametrization of the curve C, $S^1 = [0, 1]/\{0, 1\}$ is the unit interval with endpoints identified, $i : S^1 \times \mathbb{R}^+ \to \mathbb{R}$, and $\mathcal{R} : S^1 \times \mathbb{R}^+ \to \mathbb{R}^2$. The

image-based force is the term $i\mathcal{N}$, where $\mathcal{N}$ is the unit normal vector to the curve C. The regularizing and topology preserving term of the active contour is $\mathcal{R}$. We make the assumption that i is uniformly bounded, i.e.,

$$\sup_{t\in\mathbb{R}^+, p\in S^1} |i(p,t)| < +\infty. \tag{19}$$

Note that our assumptions in (17) and (19) are not too restrictive, and are typically true in practical applications. For example, the data terms of Mumford-Shah [72], Chan-Vese [17], and other region based curve evolutions satisfy these conditions when the image is bounded, which is true for digital images.

Our approach to derive R_k and $\mathcal{R}$ so that (16) and (18) preserve the topology of the initial configurations is to minimize an energy. The motion to minimize the energy as fast as possible with respect some given inner product will correspond to the terms R_k and $\mathcal{R}$. The energy approaches infinity as the curve approaches self intersection, and becomes large when the curve is "irregular". The intuition for our approach is from electrostatics. We imagine that the curve has a uniformly distributed charge along its perimeter, and that the curve moves in response to the charge as well as its original image based force. We expect that as the curve becomes close to self intersection, a repulsive force will arise due to its charge distribution and prevent self intersection. With this intuition, we propose to find the flow that minimizes the following energy functional

$$E_\gamma(C) = \frac{1}{2}\iint_{C\times C} \frac{\mathrm{d}\hat{s}\,\mathrm{d}s}{\|C(\hat{s}) - C(s)\|^\gamma} \tag{20}$$

where C denotes a curve, $\mathrm{d}s$ and $\mathrm{d}\hat{s}$ are arclength measures, $\|\cdot\|$ is the usual Euclidean norm, and $\gamma > 0$. Each pair of points on the curve contributes an amount inversely proportional to its distance to the total energy. Note that in the case when $\gamma = 1$, E_γ is the electric potential energy of the charge configuration assuming a three dimensional flux.

It should be noted that for $\gamma \geq 1$, $E_\gamma = +\infty$ for all curves. In this case, the inner integral diverges for every $C(s)$. If we consider a small neighborhood around the point $C(s)$, then the curve can be approximated by a straight line and the inner integral around the neighborhood is approximated by $\int_{-\epsilon}^{\epsilon} \frac{\mathrm{d}x}{|x|^\gamma}$ where $\epsilon > 0$ is a small number related to the size of the neighborhood around $C(s)$. This integral diverges in the case when $\gamma \geq 1$, and thus so does E_γ. In the case when $\gamma < 1$, the energy is finite (as E_γ converges if and only if $\int_{-\epsilon}^{\epsilon} \frac{\mathrm{d}x}{|x|^\gamma} < \infty$); however, one can show that as a point on the curve approaches a segment of a curve that the gradient of the energy is finite. That is, for $\gamma < 1$, a topology change, i.e., self intersection, can occur if there is an external image-based term present.

Energies similar to E_γ have been considered in mathematical literature on knot energies [76, 33, 77, 1]. A knot is an embedding of the circle S^1 into a three dimensional manifold that is usually taken to be $\mathbb{R}^3$. Two knots are said to be equivalent (or have the same knot type) if there exists an orientation preserving homeomorphism of $\mathbb{R}^3$ that maps one knot onto the other. As knots can be very complicated, it is of interest in the mathematical literature to introduce invariants of knots. One approach by O'Hara [76, 77] proposes an energy that becomes infinite as the knot changes its knot type, i.e., it begins to self intersect. An energy proposed is

$$E_\gamma(C) = \frac{1}{2} \iint_{C \times C} \left(\frac{1}{\|C(\hat{s}) - C(s)\|^\gamma} - \frac{1}{d_C(\hat{s}, s)^\gamma} \right) \mathrm{d}\hat{s}\, \mathrm{d}s \tag{21}$$

where $1 < \gamma < 3$, and $d_C(s, \hat{s})$ is the geodesic distance along the curve C from point $C(s)$ to point $C(\hat{s})$. The geodesic distance is the shortest distance along the curve from one point to another. The second term of (21) is used to "subtract out" the infinity from the first term, and thus the result is that the energy is finite. However, the property that the energy becomes infinite as the knot changes knot type still holds. Our approach will be to consider (21) with $\gamma = 1$ for active contours, which has not been considered for use as a knot invariant in the mathematical literature because it does not diverge as the curve changes knot-type. We choose the energy with $\gamma = 1$

since, unlike the energy for $\gamma \geq 2$, it roughly favors curves with shorter length. This can be seen by evaluating the energy of a circle with radius R as a function of γ. Indeed, for $\gamma \geq 2$, the circle's radius does not favor smaller radii circles. Therefore, it seemed the corresponding gradient flow for $\gamma = 1$ would have a length shrinking effect, that would regularize the curve as in the curve shortening flow; indeed this is the case as we will show. Moreover, based on experimental and analytical results, we believe the gradient flow still preserves topology.

2.2.1 Polygon Case

In this section, we define an energy for the case of polygons, and then derive its gradient flow. We define an n-polygon, where $n \geq 3$, to be a closed contour, embedded in the plane, consisting of a n straight line segments. First, we introduce some notation: $\mathbb{Z}_n = \{0, 1, \ldots, n-1\}$, C is the set of all points in the polygon, v_i is the i^{th} vertex of C where $i \in \mathbb{Z}_n$, and C_i is the edge of C connecting v_i to v_{i+1}.

2.2.1.1 Energy

We define an energy on n-polygons, $E_p : \underbrace{\mathbb{R}^2 \times \cdots \times \mathbb{R}^2}_{n \text{ times}} \to \mathbb{R}$, as

$$E_p(v_0, \ldots, v_{n-1}) = \sum_{i \in \mathbb{Z}_n} \left(|C_i| \ln |C_i| - |C_i| \right) + \frac{1}{2} \sum_{i \neq j \in \mathbb{Z}_n} \iint_{C_i \times C_j} \frac{\mathrm{d}\hat{s}\, \mathrm{d}s}{\|C_i(s) - C_j(\hat{s})\|} \quad (22)$$

where $|C_i|$ denotes the length of C_i. First, we note that the terms where the integral in (20) diverges, that is, the self-energies, $\iint_{C_i \times C_i} \frac{\mathrm{d}\hat{s}\, \mathrm{d}s}{\|C(\hat{s}) - C(s)\|}$, are replaced by the first term of (22). It should be noted that the first term in (22) arises from taking the "finite part" of the integral $\iint_{C_i \times C_i} \frac{\mathrm{d}\hat{s}\, \mathrm{d}s}{\|C(\hat{s}) - C(s)\|}$, that is the self-energies, and discarding the "infinite component". This is somewhat similar to the ideas in [76, 77] where the infinite part of the integral is subtracted out with another component. For the specific case of polygons, the self-energies can be computed explicitly, in an analytic form, and the "infinite component" can be removed, leaving the "finite component".

Therefore, the first term is simply the self energies with the infinite component removed. Further, we note that the energy defined above is E_γ in (20) for $\gamma = 1$, with the modification mentioned to keep the energy finite; it is not the energy in (21) evaluated for a polygon. The second term in (22) is the energy in (20) with $\gamma = 1$ excluding the interactions between a segment and itself, that is, excluding the self-energies. Second, we remark that the polygon energy in (22) has the desirable property that as the polygon approaches self intersection, the gradient of energy becomes infinite. We believe that adding the resulting gradient descent flow of (22) to a image-based evolution will prevent self intersection and thus, a topology change. An analytical justification of this conjecture is given in Section 2.2.1.3.

A polygon regularizer using similar ideas to our approach is considered by Unal et al. in [108]. Unal *et al.* assume that the polygon consists of a uniform charge distributed along the polygon's perimeter. However, an energy is not minimized; instead, the electrostatic force at each vertex is computed numerically, and the vertex is moved in the direction of the force. The electrostatic force at each vertex is infinite due to the near neighbor effects of the adjacent segments to the vertex. To deal with the numerical problems associated with the computation of the electrostatic force, the near neighbor effects are simply ignored. While this method may prevent some topology changes, it will not prevent the angle between two adjacent sides from becoming zero; therefore it will not prevent the two adjacent sides from touching. Moreover, as the number of vertices increases and the length of the segments decreases, thus approaching the continuum, the flow appears to be unstable. We shall see that the proposed minimization of the energy in (22) does not have the problems discussed above.

2.2.1.2 Vertex Evolution

We now derive a motion of the vertexes that minimizes the energy given in (22). The energy, E_p, is defined on $\mathbb{R}^{2\times n}$ and hence the gradient of E_p is defined with respect to the usual inner product on $\mathbb{R}^{2\times n}$. Therefore, the gradient of E_p, is the vector $\nabla E_p(v_0,\ldots,v_{n-1})=(\partial E_p/\partial v_k(v_0,\ldots,v_{n-1}))_{k=0}^{n-1}$. It can be shown that

$$F_k = F_k^{\text{self},k-1} + F_k^{\text{self},k} + \sum_{i\in\mathbb{Z}_n\setminus\{k-1\}} F_k^{k-1,i} + \sum_{i\in\mathbb{Z}_n\setminus\{k\}} F_k^{k,i} \tag{23}$$

where $F_k := -\partial E_p/\partial v_k$,

$$F_i^{j,k} = -\frac{1}{2}\frac{\partial}{\partial v_i}\iint_{C_j\times C_k} \frac{\mathrm{d}\hat{s}\,\mathrm{d}s}{\|C(\hat{s})-C(s)\|},$$

and

$$F_i^{\text{self},j} = -\ln|C_j|\frac{v_i - v_j^*}{|C_j|} \text{ for } j = i-1, i. \tag{24}$$

Note that $v_j^* = v_j$ when $j = i-1$, and $v_j^* = v_{j+1}$ when $j = i$. The F's in the above expressions represent "forces"; they get their name from the fact that they are the negative gradient of some term of the energy, E_p. It is important to note, however, that F_k is not the same as the electrostatic force, which is the negative gradient of the *potential integral*, i.e., a single integral. Moreover, F_k is finite for embedded polygons, but the electrostatic force is infinite along the polygon. We have derived simplifications for F_k, which are found in Appendix A.1. The motion of the vertices to minimize E_p is

$$\frac{\mathrm{d}v_k}{\mathrm{d}t} = F_k = -\frac{\partial E_p}{\partial v_k} \quad \text{for } k\in\mathbb{Z}_n \tag{25}$$

where t denotes an artificial time variable.

2.2.1.3 Embeddedness of Polygon

In this section, we analytically justify that an active polygon with the regularizer $R_k = F_k$ does not change topology, i.e., stays embedded. It can be shown that E_p does not become infinite, in general, as a polygon approaches a topology change;

therefore, our approach is to show that R_k becomes infinite and moves the polygon away from changing topology. The system of ODE (25) is nonlinear and autonomous, and therefore, an approach to prove existence and uniqueness of the equation is to show that $R = (R_k)_{k=0}^{n-1}$ is Lipschitz. The force R is not Lipschitz on bounded sets in $\mathbb{R}^{n\times 2}$ because R is infinite for non-embedded configurations of the vertices. It may possible to show short-time existence and uniqueness of (25), then show that the vertices stay suitably far from non-embedded configurations, and then continue the solution; we are not certain of the details. Therefore, we only show that

Proposition 2.2.1 (Embeddedness of Polygon). *Suppose* $\{v_0^0, v_1^0, \ldots, v_{n-1}^0\}$ *defines an embedded* n*-polygon,* I_k *satisfies the condition in* (17)*, and a continous solution of* (16) *exists and is unique. Then the* n*-polygon evolving in time according to* (16) *with initial conditions* $v_k(0) = v_k^0$ *for* $k \in \mathbb{Z}_n$ *maintains its initial topology for all* $t \in \mathbb{R}^+$ *and any* $\alpha > 0$*.*

Proof. We assume that the n-polygon defined according to (16) changes topology in finite time and derive a contradiction. The contradiction will arise, as we shall show, from the fact that $\|R_k\| = \|F_k\| \to +\infty$ for an appropriate $k \in \mathbb{Z}_n$ and that the direction of this R_k is opposite to the direction of topology change. Since I_k remains bounded by assumption, this is enough to derive the contradiction. There are three types of topology changes as illustrated in Fig. 1.

We consider the first type of topology change, that is, we assume that the length of a segment approaches zero, or equivalently, we assume that a vertex approaches an adjacent vertex[1]. Let us assume that $|C_k| \to 0$. Without loss of generality we assume that $|C_k| < 1$. Then we see from (24) that $\|F_k^{\text{self},k}\| = -\ln|C_k|$ since $|C_k| = \|v_k - v_{k+1}\|$, and thus $\lim_{|C_k|\to 0} \|F_k^{\text{self},k}\| = +\infty$. Moreover, we see that $F_k^{\text{self},k}$ is in the direction of $v_k - v_{k+1}$, which moves v_k away from v_{k+1}. Thus, it is impossible for

[1]Note that this case is not a topology change for $n \neq 3$ in the usual sense; however, we refer to this as a topology change since the polygon does not remain an n-polygon.

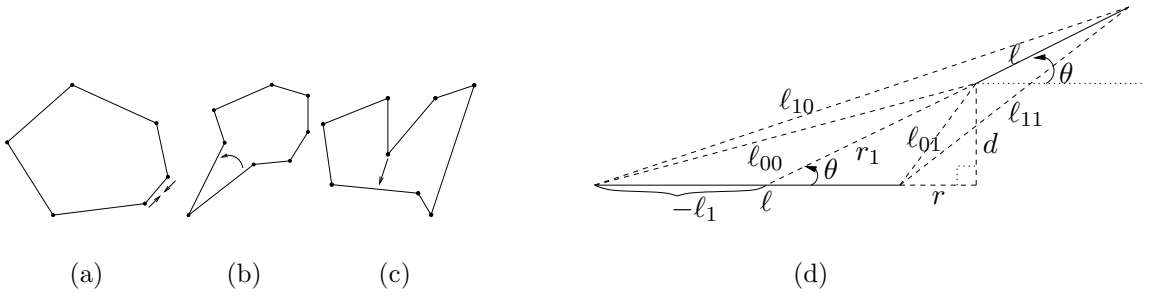


Figure 1: Illustration of three different cases of polygon topology change or self intersection are shown in (a), (b), (c). Definition of quantities used in justifying Proposition 2.2.1 is given in (d).

$|C_k| \to 0$.

Now we assume that the angle between two adjacent segments, C_k and C_{k-1}, approaches zero. By the previous case, we can assume that the lengths of the two adjacent segments do not approach zero. Further, we assume that both segments have the same length, ℓ, which remains constant with respect to the angle. If this is not the case, then because $|C_k|, |C_{k-1}| \not\to 0$ there exists segments of C_k and C_{k-1} that are adjacent and have constant lengths throughout the evolution. In this case replace C_k and C_{k-1} with these segments for the next computation. From (127) in Appendix A.1, and noting $\ell^{22}_{k-1,k} = \ell\sqrt{2}\sqrt{1-\cos\theta}$ by using the law of cosines we have

$$E_p^{k-1,k} := \iint_{C_{k-1}\times C_k} \frac{\mathrm{d}\hat{s}\,\mathrm{d}s}{\|C(\hat{s}) - C(s)\|} = 2\ell \ln\left(1 + \frac{\sqrt{2}}{\sqrt{1-\cos\theta}}\right) \tag{26}$$

where θ is the angle between the segments C_k and C_{k-1}. It is clear that $E_p^{k-1,k} \to +\infty$ as $\theta \to 0^+$. Hence it follows that $\lim_{\theta\to 0} \partial E_p^{k-1,k}/\partial\theta = -\infty$. Therefore, $\|F_k\| \to +\infty$, and F_k is in the direction to increase θ. This contradicts the assumption that $\theta \to 0$.

Next, assume a vertex, v_k, of the n-polygon approaches a segment, C_l, which is not adjacent to the segment containing v_k and does not contain v_k. We do not lose generality by assuming that $\|v_k - v_l\| \to 0$ where v_l is a vertex of C_l. If this is not the case, then clearly there is some point $p \in C_l$ such that $\|p - v_k\| \to 0$. In this case, replace v_l by p and ignore a segment of C_l that is on one side of p. Just as

in the previous case, we may assume that $|C_l| = |C_k| = \ell$ without loss of generality. The diagram in Fig. 1(d) illustrates v_k and C_l, and labels some quantities that will be used. In this diagram, we can assume $\theta \in [0, \pi)$. For the moment we assume that $\theta \neq 0$ for all time up to and including the time of topology change. After an integration, we see that

$$E_p^{k,l} := \iint_{C_l \times C_k} \frac{\mathrm{d}\hat{s}\ \mathrm{d}s}{\|C(\hat{s}) - C(s)\|} = \sum_{i,j=0}^{1} (-1)^{j+1} \ell_{ij}^* \ln \left(\frac{\ell_{jj} + \ell_{\overline{i \oplus j}, i \oplus j} + \ell}{\ell_{jj} + \ell_{\overline{i \oplus j}, i \oplus j} - \ell} \right) \tag{27}$$

where $\oplus$ denotes the xor operation, $\overline{\cdot \oplus \cdot}$ denotes not xor, the quantities on the right hand side are defined in Fig. 1(d), $\ell_{00}^* = \ell_1$, $\ell_{01}^* = \ell + \ell_1$, $\ell_{10}^* = r_1$, and $\ell_{11}^* = \ell + r_1$. Note that the independent variables are d, r and θ. Differentiating (27) with respect to d we find

$$\frac{\partial E_p^{k,l}}{\partial d} = \sum_{i,j=0}^{1} (-1)^{j+1} \left[\frac{\partial \ell_{ij}^*}{\partial d} \ln \left(\frac{\ell_{jj} + \ell_{\overline{i \oplus j}, i \oplus j} + \ell}{\ell_{jj} + \ell_{\overline{i \oplus j}, i \oplus j} - \ell} \right) - \frac{2\ell \ell_{ij}^*}{(\ell_{jj} + \ell_{\overline{i \oplus j}, i \oplus j})^2 - \ell^2} \frac{\partial}{\partial d} (\ell_{jj} + \ell_{\overline{i \oplus j}, i \oplus j}) \right]. \tag{28}$$

It is clear that the terms of the sum in (28) for $(i, j) \neq (1, 0)$ have finite limit as $(d, r) \to (0, 0)$. Thus, to show $\partial E_p^{k,l}/\partial d \to -\infty$, we only need to show the term with $(i, j) = (1, 0)$ diverges. To show this, first note that $\ell_{00} = \sqrt{(r + \ell)^2 + d^2}$, $\ell_{01} = \sqrt{r^2 + d^2}$, and $\ell_{10}^* = d/\sin\theta$. This gives that $(\ell_{00})_d = d/\ell_{00}$, $(\ell_{01})_d = d/\ell_{01}$ and $(\ell_{10})_d^* = 1/\sin\theta$ where the subscript d denotes partial differentiation. With some manipulation, we find that the second term in (28) has limit

$$\lim_{(r,d) \to (0,0)} \frac{\ell}{\ell_{00} \sin\theta} \frac{d^2(\ell_{01} + \ell_{00})}{\ell_{01}(r^2 + r\ell + d^2) + (r^2 + d^2)\ell_{00}} = \frac{\ell}{\sin\theta} \lim_{(r,d) \to (0,0)} \frac{d^2}{\ell_{01}(r^2 + r\ell + d^2) + (r^2 + d^2)\ell_{00}}. \tag{29}$$

We claim that (29) has finite limit along any path. To show this, note that locally around $(r, d) = (0, 0)$, a path may be represented as a function. Further, from continuity of the flow, the function is continuous. There are two cases. First, $r = r(d)$

and $r'(0)$ exists in that case, $r = \mathcal{O}(d)$, where $\mathcal{O}$ denotes big-oh. It is easy to verify that the limit in (29) is finite in this case. The second case is when $d = d(r)$ and $d'(0)$ exists in that case, $d = \mathcal{O}(r)$. Again, it is easy to verify that (29) is finite. Therefore, we see that

$$\lim_{(r,d)\to(0,0)} \frac{\partial E_p^{k,l}}{\partial d} = c - \lim_{(r,d)\to(0,0)} \frac{1}{\sin\theta} \ln\left(\frac{\ell_{00}+\ell_{01}+\ell}{\ell_{00}+\ell_{01}-\ell}\right) = -\infty. \tag{30}$$

Hence, it is clear that $\|F_k\| \to +\infty$, and the direction of F_k is in a direction away from $d \to 0$. This contradicts the assumption that $d \to 0$. In the case when $\theta = 0$, none of the above equations hold. However, since it is intuitive that $\partial E_p^{k,l}/\partial d$ is continuous in the angle θ, we may approximate $\partial E_p^{k,l}/\partial d$ at $\theta = 0$ by $\theta = \epsilon$ for ϵ small. Therefore, in this case too, we obtain a contradiction. □

2.2.2 Differentiable Contour Case

In this section, we shall consider the case when the curve is not restricted to be a polygon; the curve is allowed to be any twice differentiable curve.

2.2.2.1 Energy Functional

Let $\mathcal{C}$ denote the set of all twice differentiable embedded curves in $\mathbb{R}^2$. Then we define an energy $E_c : \mathcal{C} \to \mathbb{R}^+$ as follows

$$E_c(C) = \frac{1}{2} \iint_{C\times C} \left(\frac{1}{\|C(s) - C(\hat{s})\|} - \frac{1}{d_C(s,\hat{s})}\right) \mathrm{d}\hat{s}\,\mathrm{d}s \tag{31}$$

where $\|\cdot\|$ is the usual Euclidean norm, s and $\hat{s}$ denote arclength parameters, and $d_c(\cdot,\cdot)$ denotes geodesic distance. Note that we set $d_C(s,\hat{s}) = +\infty$ if there is no path along C connecting $C(s)$ to $C(\hat{s})$, which is possible if the curve consists of two or more closed curves. First, note that the energy in (31) is the same energy as in (21) with $\gamma = 1$, and therefore it is not a knot energy, i.e., it does not, in general, diverge as a curve approaches self-intersection. Nevertheless, we shall give analytical jjustification in Section 2.2.2.4 that the gradient flow will preserve the embeddedness of a curve.

Second, the proof that the integral in (31) exists and is not infinite for embedded curves is given in [77]. Third, it is important to note that the regularization term, i.e., the second term, in (31) only "cancels out" the infinity of the electrostatic term, i.e., first term, due to points on the curve that are close together in the geodesic sense. The regularization does not affect the asymptotic behavior of the electrostatic term as a curve approaches self-intersection. This is because points on the curve that touch during self-intersection are far away in the geodesic sense. Finally, we note that the integral of the regularization term does not depend on the geometry of the curve; it only depends on the length of the curve. In fact, one can think of the energy E_c of a curve, formally, as the difference of the electrostatic energy, which is the energy in (20) with $\gamma = 1$, and twice the electrostatic energy of a line segment that is half of the length of the curve.

2.2.2.2 Curve Evolution

For each curve $C \in \mathcal{C}$, we use the following inner product that is defined on the set of perturbations of a curve:

$$\left\langle \tilde{C}_1, \tilde{C}_2 \right\rangle_C = \int_{S^1} \tilde{C}_1(p) \cdot \tilde{C}_2(p) \|C'(p)\| \, \mathrm{d}p, \tag{32}$$

where $S^1 = [0,1]/\{0,1\}$, $\mathrm{d}p$ denotes the push forward of Lebesgue measure onto S^1, $\tilde{C}_1, \tilde{C}_2 \in C^2(S^1, \mathbb{R}^2)$ are perturbations of C, and $C \in C^2(S^1, \mathbb{R}^2)$ represents a parametrization of the curve C. The goal of this section is to find the perturbation of C that decreases the energy E_c the fastest with respect to the inner product defined in (32). By the Cauchy-Schwartz inequality, this perturbation is the negative gradient (if it exists), $-\nabla E_c(C)$, which is defined by the relation

$$\mathrm{d}E_c(C) \cdot \tilde{C} = \left\langle \tilde{C}, \nabla E_c(C) \right\rangle_C \tag{33}$$

where $\mathrm{d}E_c(C) \cdot \tilde{C}$ denotes the variation of E_c at C in the direction of a perturbation $\tilde{C}$. With this definition, we state the following proposition.

Proposition 2.2.2 (Gradient of E_c). *The gradient of E_c, defined in* (33)*, is given by*

$$-\nabla E_c(C)(s) = \lim_{\epsilon \to 0^+} \left[\mathcal{E}_\epsilon(s) + \mathcal{P}_\epsilon(s)\kappa(s) - \ln\left(\frac{L}{2\epsilon}\right)\kappa(s) \right] \mathcal{N}(s). \tag{34}$$

where

$$\mathcal{E}_\epsilon(s) = \int_{B_C(\epsilon,s)} \frac{C(s) - C(\hat{s})}{\|C(s) - C(\hat{s})\|^3} \cdot \mathcal{N}(s)\, \mathrm{d}\hat{s} \text{ and } \mathcal{P}_\epsilon(s) = \int_{B_C(\epsilon,s)} \frac{\mathrm{d}\hat{s}}{\|C(s) - C(\hat{s})\|}, \tag{35}$$

κ denotes the curvature of C, $\mathcal{N}$ is the unit inward normal to C, L denotes the length of C, and $B_C(\epsilon, s) = \{C(\hat{s}) : d_C(\hat{s}, s) > \epsilon\}$.

Proof. The computation is quite long. We refer the interested reader to idea in Appendix A.2. □

Therefore, the curve evolution given by the PDE,

$$C_t = -\nabla E_c(C) =: \mathcal{R} \tag{36}$$

minimizes the energy E_c optimally, in the sense of the inner product given in (32). Notice that this flow at a particular point depends on all other points on the curve in the form of integrals around the curve. This evolution therefore satisfies the property, mentioned in the Introduction, that each point of the curve must have some "knowledge" of other points on the curve in order to preserve topology in the presence of bounded image-based forces.

The term $\mathcal{E}_\epsilon(s) + \mathcal{P}_\epsilon(s)\kappa(s)$ arises from the electrostatic term, or the first term, of E_c, and the term $\ln(L/2\epsilon)\kappa(s)$ arises from the regularization term, or second term, of E_c. It should be noted that although $\mathcal{R}$ appears to have a term that is negative curvature, when the limit is evaluated, there is only a positive curvature component, and thus the flow is stable. For the sake of gaining intuition, let us ignore the regularization term (the term that keeps the quantities finite) in both E_c and $\mathcal{R}$. The energy $E_c(C)$ can then be regarded as the potential energy of a uniform charge distributed along the curve C, as alluded to at the beginning of Section 2.2.

The term, $\mathcal{E}_\epsilon(s)$, can be regarded as the projection of the electric field vector of the charge distribution at the point $C(s)$ onto the inward normal of C. The term, $\mathcal{P}_\epsilon(s)$, can be regarded as the electrostatic potential of the charge distribution at the point $C(s)$. With these observations, we verify that the motion of the curve under the flow, $C_t = \mathcal{R}$, corresponds to a reduction of the energy E_c. Notice that there are two factors that contribute to the energy E_c. One factor is the length of the curve, which is equivalent to total charge because of our assumption of uniform charge distribution. The other factor is the "closeness" of points on the curve to other points on the curve. Thus, to reduce energy, the curve should contract, which is equivalent to reducing its charge, in such a way that is consistent with close points moving away from each other due to repulsion of charge. Notice that the term, $\kappa\mathcal{P}_\epsilon\mathcal{N}$ of the flow given in (36), corresponds to contraction of the curve since $\mathcal{P}_\epsilon$ is positive, and any positive function multiplied by the curvature vector shrinks the length of the evolving curve. Moreover, since this term is a curvature flow, it keeps the curve smooth and away from irregularities. Repulsion from close points arises from the term, $\mathcal{E}_\epsilon\mathcal{N}$, which is the electrostatic force. The intuition given above applies also when the regularization terms are considered in the flow. The purpose of the regularization term is to cancel out the infinity arising from $\mathcal{E}_\epsilon + \mathcal{P}_\epsilon\kappa$ as $\epsilon \to 0$.

2.2.2.3 Behavior of the Circle

We consider a simple, yet insightful, example of a circle whose energy, E_c, and negative gradient, $\mathcal{R}$ are to be computed. Let us assume that the circle has a radius of length R, and it is centered at the origin of $\mathbb{R}^2$. Note that by symmetry of the circle, it is clear that $\mathcal{E}_\epsilon(s)$, $\mathcal{P}_\epsilon(s)$, and the inner integral of E_c do not depend on s. Therefore, we choose for simplicity to evaluate the previous quantities at $s = 0$, i.e., the point $(R, 0)$. We use the parametrization $C(\theta) = R(\cos\theta, \sin\theta)$ where $\theta \in [0, 2\pi]$ and $\hat{s} = R\theta$. A

computation gives

$$\|C(0) - C(\hat{s})\| = 2R \sin \frac{\theta}{2} \text{ for } \theta \in [0, 2\pi] \quad \text{and} \quad d_C(0, \hat{s}) = R\theta \text{ for } \theta \in [0, \pi]. \quad (37)$$

Now, we see for the circle that

$$\begin{aligned} E_c(C) &= \lim_{\epsilon \to 0^+} \pi R \int_{B_C(\epsilon, 0)} \left(\frac{1}{\|C(0) - C(\hat{s})\|} - \frac{1}{d_C(0, \hat{s})} \right) \mathrm{d}\hat{s} \\ &= 2\pi R \lim_{\epsilon \to 0^+} \int_{\frac{\epsilon}{R}}^{\pi} \left(\frac{1}{2R \sin \frac{\theta}{2}} - \frac{1}{R\theta} \right) R \, \mathrm{d}\theta \end{aligned} \quad (38)$$

where we have substituted (37) into (38). By standard calculus techniques, we find

$$\begin{aligned} E_c(C) &= 2\pi R \lim_{\epsilon \to 0+} \ln \left(\frac{\frac{\epsilon}{R} \sin \epsilon/2R}{\pi(1 - \cos \epsilon/2R)} \right) \\ &= 2\pi R \lim_{\epsilon \to 0^+} \ln \left(\frac{\epsilon^2/2R^2}{\frac{\pi}{2}\epsilon^2/(2R)^2} \right) = 2\pi R \ln \frac{4}{\pi}. \end{aligned} \quad (39)$$

Noting that the inward normal is given by $\mathcal{N}(0) = (-1, 0)$, and that $(C(0) - C(\hat{s})) \cdot \mathcal{N}(0) = -R(1 - \cos\theta) = -2R \sin^2 \theta/2$, we see that

$$\begin{aligned} \mathcal{E}_\epsilon(0) &= \int_{B_C(\epsilon, 0)} \frac{C(0) - C(\hat{s})}{\|C(0) - C(\hat{s})\|^3} \cdot \mathcal{N}(0) \, \mathrm{d}\hat{s} \\ &= -\frac{1}{2R} \int_{\frac{\epsilon}{R}}^{2\pi - \frac{\epsilon}{R}} \frac{\mathrm{d}\theta}{2 \sin \frac{\theta}{2}} = -\frac{1}{2R} \ln \left(\frac{1 + \cos \epsilon/2R}{1 - \cos \epsilon/2R} \right). \end{aligned}$$

Using a similar computation as for E_c, the expression

$$\mathcal{P}_\epsilon(0) = 2 \ln \left(\frac{\sin \epsilon/2R}{1 - \cos \epsilon/2R} \right) \quad (40)$$

is easily derived. Clearly, the curvature satisfies $\kappa(0) = 1/R$ for the circle. Hence, we see

$$\begin{aligned} \mathcal{R}(0) = \lim_{\epsilon \to 0^+} \Big[-\frac{1}{2R} \ln \left(\frac{1 + \cos \epsilon/2R}{1 - \cos \epsilon/2R} \right) + \frac{2}{R} \ln \left(\frac{\sin \epsilon/2R}{1 - \cos \epsilon/2R} \right) - \frac{1}{R} \ln \left(\frac{\pi R}{\epsilon} \right) \Big] \\ = \frac{1}{R} \ln \frac{4}{\pi} \end{aligned} \quad (41)$$

where standard calculus techniques were applied.

First observe that the energy computed in (39) clearly reconciles with the flow $C_t(s) = \ln(4/\pi)/R\mathcal{N}(s)$ computed in (41). Second, we see from the flow that the

circle shrinks to a point, in fact, this happens in finite time. This property is also seen in the curve shortening flow. This suggests that the flow $C_t = \mathcal{R}$ may shrink arbitrary curves to points in finite time. These remarks clearly show similarities between curve shortening and the flow in (36).

2.2.2.4 Embeddedness of Curve

Although there are similarities between curve shortening flow and the flow of (36), in this section we highlight a major difference. Namely, curve shortening flow does not preserve the topology of a curve when combined with a bounded image-based force, although it preserves topology in the absence of another force. Since we are not aware of any existence or uniqueness results, or any results that say that the curvature stays well-defined for all time as the curve evolves according to the integral PDE (36), we can only give a non-rigorous, intuitive justification that an active contour evolving with (36) and a bounded force stays embedded. We can only justify is that if the curve stays smooth, then as the curve approaches topology change, $\mathcal{R}$ becomes infinite and points in a direction away from topology change.

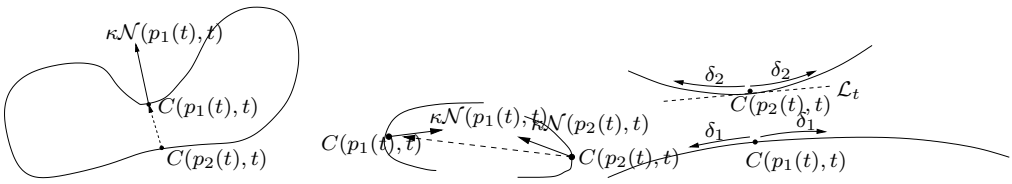
Conjecture 2.2.1 (Embeddedness of Curve). *Suppose $C_0 \in C^2(S^1, \mathbb{R})$ defines an embedded curve and $i \in C^2(S^1 \times \mathbb{R}^+, \mathbb{R})$ satisfies the condition in* (19)*. Then the curve evolving in time according to* (18) *with initial condition $C(\cdot, 0) = C_0$ stays embedded for all $t \in \mathbb{R}^+$ for any $\alpha > 0$.*

Proof. We sketch our justification in several steps. Let us first define for convenience,

$$I(p,q,t) := \left(\frac{\kappa(p,t)}{\|C(p,t) - C(q,t)\|} + \frac{C(p,t) - C(q,t)}{\|C(p,t) - C(q,t)\|^3} \cdot \mathcal{N}(p,t) \right) \|C_p(q,t)\| \mathcal{N}(p,t).$$

1. We assume for the sake of contradiction that $C(\cdot, t)$ changes topology, for the first time, in finite time $t_i \in \mathbb{R}^+$. That is, $\exists\ \delta_0 > 0$ and $p_1, p_2 \in C([0, t_i]; S^1)$ such that

$$\lim_{t \to t_i} \|C(p_1(t), t) - C(p_2(t), t)\| = 0 \text{ and } d_{C(\cdot,\cdot,t)}(p_1(t), p_2(t)) > \delta_0 \forall t \in [0, t_i).$$



(a) (b) (c)

Figure 2: A curve moving toward topology change with quantities labeled as in the justification of Conjecture 2.2.1 is shown in (a). A topology change cannot occur, that is $C(p_2(t),t) \nrightarrow C(p_1(t),t)$, if locally, the curve and directions are indicated in (b) for all time close to t_i. (c) shows a diagram for quantities defined in Step 4 and 5 of the proof.

2. By translating the coordinate axis so that $C(p_1(t),t)$ coincides with the origin, we may assume that $\mathrm{d}/\,\mathrm{d}tC(p_1(t),t) = 0$. This, along with the conclusion in Step 1, implies that $\exists t_0 \in (0,t_i)$ such that $\mathrm{d}/\,\mathrm{d}tC(p_2(t),t) \neq 0\ \forall t \in [t_0,t_i)$. We may now assume without loss of generality that $(\,\mathrm{d}/\,\mathrm{d}tC(p_2(t),t)) \cdot \kappa(p_1(t),t)\mathcal{N}(p_1(t),t) \geq 0\ \forall t \in [t_0,t_i)$. This statement says that $C(p_2(t),t)$, after some time, moves with a component in the same direction as $\kappa(p_1(t),t)\mathcal{N}(p_1(t),t)$. If this is not the case, then simply switch the roles of $p_1(t)$ and $p_2(t)$, and the statement will hold. We give an intuitive argument for this statement illustrated in Fig. 2(b). If the opposite of the statement is true, then for time near enough to t_i, the curve locally looks as shown in Fig. 2(b). But a topology change cannot occur in this case.

3. We must prove that the curve $C(\cdot,t)$ remains smooth, that is sufficiently differentiable, for $t \in [0,t_i)$. We do not have the details for this, but comment that this seems intuitive since the expression for $\mathcal{R}$ contains a curvature term, which should keep $C(\cdot,t)$ smooth.

4. By Lemma A.3.1 in Appendix A.3, $\exists \delta_1 > 0$ such that

$$\begin{aligned}
&\frac{\kappa(p_1(t),t)}{\|C(p_1(t),t)-C(p,t)\|} = \frac{1}{s(t)}\kappa(p_1(t),t) - \frac{1}{8}\kappa^3(p_1(t),t)s(t) + \mathcal{O}(s^2(t)) \\
&\frac{C(p_1(t),t)-C(p,t)}{\|C(p_1(t),t)-C(p,t)\|^3}\cdot \mathcal{N}(p_1(t),t) \\
&\qquad\qquad = -\frac{1}{2s(t)}\kappa(p_1(t),t) - \frac{1}{6}\kappa'(p_1(t),t) + \mathcal{O}(s(t))
\end{aligned}$$

for $p \in S^1$ satisfying $s(t) := d_{C(\cdot,t)}(p_1(t),p) < \delta_1$. We assume that δ_1 does not depend on $t \in (t_0,t_i)$. Therefore, for t fixed, the previous estimates yield

$$\begin{aligned}
\mathcal{R}_1(p_1(t),t) &:= \lim_{\epsilon\to 0^+}\left(\int_{A_\epsilon(t)} I(p_1(t),p,t)\,\mathrm{d}p - \ln\frac{L(t)}{2\epsilon}\kappa(p_1(t),t)\mathcal{N}(p_1(t),t)\right) \\
&= \lim_{\epsilon\to 0^+}\left\{2\int_\epsilon^{\delta_1}\left(\frac{1}{2s}\kappa(p_1(t),t) - \frac{1}{6}\kappa'(p_1(t),t) + \mathcal{O}(s(t))\right)\mathrm{d}s \right. \\
&\qquad \left. - \ln\frac{L(t)}{2\epsilon}\kappa(p_1(t),t)\right\}\mathcal{N}(p_1(t),t) \\
&= \left(\ln\frac{2\delta_1}{L(t)}\kappa(p_1(t),t) - \frac{1}{3}\delta_1\kappa'(p_1(t),t) + \mathcal{O}(\delta_1^2)\right)\mathcal{N}(p_1(t),t)
\end{aligned}$$

where $A_\epsilon(t) := \{p \in S^1 : \epsilon < d_{C(\cdot,t)}(p_1(t),t) < \delta_1\}$. In particular, since $L(t) \nrightarrow 0$ as $t \to t_i$ because of the result in Step 1 and $L(t) \nrightarrow +\infty$ as $t \to t_i$, we have that $\mathcal{R}_1(p_1(t),t)$ stays uniformly bounded for $t \in [t_0,t_i)$.

5. Now, since $C(\cdot,t)$ is smooth for $t \in [0,t_i)$, $\exists \delta_2 > 0$ such that $C(p,t)$ can be approximated by a line segment $\mathcal{L}_t$ passing through the point $C(p_2(t),t)$ for $p \in S^1$ satisfying $d_{C(\cdot,t)}(p_2(t),p) < \delta_2$. We assume that δ_2 is independent of $t \in [t_0,t_i)$. By Step 1, the neighborhood for which the approximation of $\mathcal{L}_t$ holds is far away from $C(p_1(t),t)$ in the geodesic sense.

6. Setting $B_t = \{p \in S^1 : d_{C(\cdot,t)}(p_1(t),p) > \delta_1, d_{C(\cdot,t)}(p_2(t),p) > \delta_2\}$ and noting that the Euclidean of $C(p_1(t),t)$ to points in the set B_t remains far away from zero for $t \in (t_0,t_i)$, we see that

$$\mathcal{R}_2(p_1(t),t) := \int_{B_t} I(p_1(t),p,t)\,\mathrm{d}p$$

is uniformly bounded over $t \in (t_0,t_s)$.

7. By Step 5 and elementary integration, we see that

$$\begin{aligned}\int_{C_t} I(p_1(t),p,t)\,\mathrm{d}p &\approx \int_{\mathcal{L}_t} \frac{\mathrm{d}s}{\|C(p_1(t),t)-\mathcal{L}_t(s)\|}\kappa(p_1(t),t)\mathcal{N}(p_1(t),t)\\ &+ \int_{\mathcal{L}_t} \frac{C(p_1(t),t)-\mathcal{L}_t(s)}{\|C(p_1(t),t)-\mathcal{L}_t(s)\|^3}\cdot\mathcal{N}(p_1(t),t)\,\mathrm{d}s\mathcal{N}(p_1(t),t)\\ &= \ln\left(\frac{r_1(t)+r_2(t)+\ell}{r_1(t)+r_2(t)-\ell}\right)\kappa(p_1(t),t)\mathcal{N}(p_1(t),t)\\ &+ \frac{2}{(r_1(t)+r_2(t))^2-\ell^2}\left(\frac{\vec{r_1}(t)}{r_1(t)}+\frac{\vec{r_2}(t)}{r_2(t)}\right)\cdot\mathcal{N}(p_1(t),t)\mathcal{N}(p_1(t),t) \qquad (42)\end{aligned}$$

where $\vec{r_1}(t)$ and $\vec{r_2}(t)$ are the vectors from the endpoints of $\mathcal{L}_t$ to $C(p_1(t),t)$ and $r_1(t)$ and $r_2(t)$ are the norms of the vectors, ℓ is the length of $\mathcal{L}_t$, which is assumed constant without loss of generality because of the independence of δ_2 with time, and $C_t := \{p \in S^1 : d_{C(\cdot,t)}(p_1(t),p) < \delta_2\}$. By Step 1, we have that $\lim_{t\to t_i}\|C(p_1(t),t)-C(p_2(t),t)\| = 0$, and since $C(p_2(t),t) \in \mathcal{L}_t$ for $t \in (t_0,t_i)$, we conclude that $r_1(t)+r_2(t) \to \ell$ as $t \to t_i$. It is clear that for t near t_i, the second term of (42) points in the same direction as $\kappa(p_1(t),t)\mathcal{N}(p_1(t),t)$. Moreover, the norm of both terms of (42) approach $+\infty$ as $t \to t_i$. Since we have chosen the coordinate system so that $C(p_1(t),t)$ remains fixed, it follows that for t close to t_i, we have that $(\,\mathrm{d}/\,\mathrm{d}tC(p_2(t),t))\cdot\kappa(p_1(t),t)\mathcal{N}(p_1(t),t) < 0$. But this contradicts the conclusion in Step 2.

□

2.2.2.5 Remarks on Related Work

There has been a recent work in the computer vision literature, namely the work of Rochery et al. [86] that uses the idea of energies that are defined as double integrals around curves rather than single integrals that are traditionally used. One purpose of these double integral energies in [86] is to impose a global shape prior on the contour, as opposed to a prior that is only local. The goal of [86] is to segment line structures in satellite images, and thus *a priori* knowledge of the shape of the contour, which is

thin elongated structures, is incorporated. The energy proposed to be minimized to enforce these elongated structures is

$$E(C) = -\int_0^1 \int_0^1 g(\|C(p) - C(p')\|)\vec{t}(p) \cdot \vec{t}(p') \, \mathrm{d}p \, \mathrm{d}p' \tag{43}$$

where p and p' are parameters denoting a parametrization of C, $\vec{t}(p) = C'(p)$ and $g : \mathbb{R} \to \mathbb{R}$ is heuristically chosen so that E will be minimized by thin elongated structures.

We now illustrate the differences between the model of Rochery et al. and our model. First, we note that the energy in (43) contains the term $\vec{t}(p) \cdot \vec{t}(p')$ in the integrand. This is different in our model in which the contribution a pair of points makes to the energy depends *only* on the distance between them in the Euclidean and geodesic sense. We learned from a reviewer that this tangent term leads to producing 'arms' of the contour, which is not desirable for the purpose of this work. Another difference between the model of Rochery and our model is the lack of the term $\kappa P\mathcal{N}$, or any other regularization term arising from the gradient flow that minimizes (43). Indeed, for regularity, the authors of [86] add a separate curvature term to the flow. Finally, (43) does not allow for a term inside the integrand depending on geodesic distance, as in our model. Indeed, the general formulation of [86] does not allow for a function inside the integrand that is also a function of the curve, C, as in the geodesic distance. Deriving the gradient flow for this type of term requires a different computation than the computation for (43).

We learned of the recent work in [87] that also considers energies of the form $\int_C \int_C \phi(C(s), C(\hat{s})) \, \mathrm{d}\hat{s} \, \mathrm{d}s$, although gradient flows of these energies are not considered, and the idea is not pursued. The first term of the energy we consider falls in the form of this energy.

2.3 Numerical Implementation

In this section, we describe possible approaches of implementing the polygon vertex evolution given in (25) and the curve evolution given in (36).

2.3.1 Polygons

We implement the polygon evolution using a standard marker particle method. That is, the vertexes of the evolving polygon are stored in a $N \times 2$ array where N is the number of vertexes of the polygon. The vertices are assumed to be ordered so that v_k is adjacent to v_{k+1} for all $k \in \mathbb{Z}_n$. Discretizing the continuous evolution given in (25) yields

$$v_k^{n+1} = v_k^n + \Delta t F_k^n \tag{44}$$

where v_k^n denotes the k^{th} vertex at time n, similarly for the force F_k^n, and Δt is the step size, which is a small positive number. There are no restrictions on the step size to guarantee stability. Typically, we use $\Delta t = 1$. However, due to the fact that finite time steps are used, it may be necessary to use an additional discrete topology preservation step to guarantee that the topology is preserved numerically (see Section 2.3.2.2 for a detailed discussion and possible scheme). Note that F_k^n is the sum of N forces that are given in (24). The final expressions for these forces are given in Appendix A.1. Note that for segments that are not adjacent, the expression for the force given in Appendix A.1 is not closed form, and thus a numerical integration must be performed. We use a simple Riemann sum approximation to approximate this integral. We remark that the continuous curve evolution can be done using marker particles (replacing F_k with $\mathcal{R}$) in this manner. One can use multiple contours, and the topology is guaranteed to be preserved.

2.3.2 Level Set Method

2.3.2.1 Approximating the Flow

We consider the flow

$$C_t(s) = (\mathcal{E}_\epsilon(s) + \kappa(s)\tilde{\mathcal{P}}_\epsilon(s))\mathcal{N}(s) \tag{45}$$

where $\tilde{\mathcal{P}}_\epsilon = \mathcal{P}_\epsilon - \ln(L/2\epsilon)$ and $\epsilon > 0$ chosen small, as an approximation of implementing (36). We should comment that for $\epsilon > 0$ small enough, $\tilde{\mathcal{P}}_\epsilon$ will be positive. Note that $-\ln(L/2\epsilon)$ counters the infinity arising from both the terms $\mathcal{E}_\epsilon$ and $\mathcal{P}_\epsilon$. In the case that ϵ is chosen quite small, the term is needed for preventing blowup of the flow due to near neighbor effects. However, when ϵ is not chosen too small, in the case that the curve is not very finely sampled, we found experimentally that ignoring the term $-\ln(L/(2\epsilon))$ better simulated the actual continuous-time behavior of the PDE. For example, in the case of the curve not being sampled (hence, $\epsilon \geq 1$), we found ignoring that term more accurately matched the behavior of the evolution when the curve was sampled very finely (hence, ϵ small) and the $-\ln L/2\epsilon$ term was included. The $-\ln(L/2\epsilon)$ mainly removes the infinity from the $\mathcal{E}_\epsilon$ and $\mathcal{P}_\epsilon$ terms as $\epsilon \to 0$, but when $\epsilon \geq 1$, the previous terms are already finite and not large, and therefore there may not be a need for using the log term, when $\epsilon \geq 1$. This may explain why ignoring the term $-\ln(L/2\epsilon)$ seems to simulate the PDE better when $\epsilon \geq 1$. Note that ignoring this term, in this case, does not prevent the flow from preventing topology change as the other terms in the flow still become infinite in the case when the curve approaches topology change.

We perform the numerical implementation of the curve evolution given in (45) using level set methods [78], although one can use a marker particle method. To implement the flow in (36), the evolving curve is embedded as the zero level set of a scalar function $\Psi : \mathbb{R}^2 \times \mathbb{R}^+ \to \mathbb{R}$ that evolves in time. The evolution of Ψ becomes

$$\Psi_t(x,t) = -\nabla\Psi(x,t) \cdot C_t(x) \quad \text{for } x \in C(t)$$

where Ψ_t denotes the derivative with respect to the second argument, and $\nabla\Psi$ denotes the gradient with respect to the first argument. Note that we have only specified the evolution of Ψ along the zero-level set of Ψ. We shall defer how to extend this evolution to all points of the domain of Ψ until later in this section. Noting that $\mathcal{N} = -\nabla\Psi/\|\nabla\Psi\|$ (if Ψ is arranged to be positive outside the curve, and negative inside the curve) the curve evolution in (36) becomes

$$\Psi_t = -\nabla\Psi \cdot \vec{E} + \kappa P\|\nabla\Psi\| \tag{46}$$

where

$$\vec{E}(s) = \int_{B_C(\epsilon,s)} \frac{C(s) - C(\hat{s})}{\|C(s) - C(\hat{s})\|^3}\,\mathrm{d}\hat{s}, \quad \text{and} P = \tilde{\mathcal{P}}_\epsilon. \tag{47}$$

We implement (46) on a finite square grid, so a discretization of the equation must be performed. Careful attention is paid to estimating the derivatives of Ψ, which are performed using appropriate differencing schemes to ensure stability and to capture the viscosity solution. We note that the first term of (46) is a transport term, and thus, an upwinding difference scheme must be used to capture the viscosity solution. The estimate of the term becomes

$$\nabla\Psi \cdot \vec{E} \approx (D_x^+\Psi)E_x\chi_{E_x>0} + (D_x^-\Psi)E_x\chi_{E_x<0} + (D_y^+\Psi)E_y\chi_{E_y>0} + (D_y^-\Psi)E_y\chi_{E_y<0}$$

where

$$D_x^+\Psi(i,j) = \Psi(i+1,j) - \Psi(i,j), \quad D_x^-\Psi(i,j) = \Psi(i,j) - \Psi(i-1,j),$$

D_y^+ and D_y^- are defined similarly, and χ denotes the indicator function. The second term of (46) can be written as

$$\kappa P\|\nabla\Psi\| = P\frac{\Psi_x^2\Psi_{yy} - 2\Psi_x\Psi_y\Psi_{xy} + \Psi_y^2\Psi_{xx}}{\Psi_x^2 + \Psi_y^2}$$

and each of the partial derivatives of Ψ are estimated using central differences. The discrete approximation to (46) using a forward Euler scheme is

$$\Psi_{n+1}(i,j) = \Psi_n(i,j) + \Delta t(-\nabla\Psi_n(i,j) \cdot \vec{E} + \kappa P\|\nabla\Psi\|_n(i,j)) \tag{48}$$

where the quantities on the right denote the discrete approximation. The time step, Δt, is chosen to ensure stability of (48), that is,

$$\Delta t < \min\left\{ \frac{1}{\max_s |E_x(s)|}, \frac{1}{\max_s |E_y(s)|}, \frac{1}{2\max_s |P(s)|} \right\}.$$

To efficiently implement (48), we note that Ψ needs to be updated only around a small neighborhood of the zero-level set of Ψ, which is called the "narrow band". We can now define $\vec{E}$ and P in the narrow band by simply defining them to be the same value as at its closest point to the contour, i.e., the zero-level set. Since the narrow band will be thin, this definition makes sense, that is, there is a unique closest point to the contour. A high-level description of the algorithm to implement the level set evolution is given below.

1. Find polygonal estimate of zero-level set of Ψ
2. Compute $\vec{E}$ and P at each vertex of polygonal estimate
3. Extend $\vec{E}$ and P to narrow band of zero-level set of Ψ
4. Update Ψ based on (48)
5. Repeat above steps until Ψ converges

2.3.2.2 Ensuring Discrete Topology Preservation

Note that because of the finite grid size and a discrete time step in the discretization (discussed in the previous section), it is possible that when the time step is not chosen sufficiently small, there may be a topology change in the discretization of the proposed PDE when there large image-based forces. Instead of trying to derive a bound on the step size to guarantee that the topology is preserved, which may lead to excessively small time steps and therefore a time consuming implementation, we *pair* the discrete topology preserving scheme of Han et al. [45, 44] (any other discrete topology scheme will suffice) with the constructed geometric flow. One may wonder what the necessity

of using the proposed flow is when the method of Han et al. already preserves discrete topology. Again, as mentioned in the introduction, topology is a very weak property of a contour and often one wants more than simple topology preservation as will be demonstrated in the experiments in Section 2.4. Two important properties that our flow provides are global regularization and gradually adjusting the flow in a continous manner so that the topology preservation usually takes place well before the curve is about to self-intersect. This makes it much less likely that the geometric pathologies of using the method of Han et al. alone will occur. We add that the method of Han et al. is quite efficient and does not add a significant computational cost to the algorithm.

Using a discrete topology preserving step is also a better simulation of the underlying PDE than not using such a step considering the discrete time step and the finite resolution grid. This is because without such a discrete step, a topology change may occur, and such a change is inconsistant with the topology preserving property of the PDE. Also, as the grid resolution becomes finer and the step size becomes smaller, then eventually the proposed method will not need to invoke the discrete topology preserving step to guaruntee topology preservation, and the evolution with the discrete topology step approaches the true solution of the PDE in this case. Therefore, the use of the discrete step leads to a good approximation of the true solution of the PDE, which is also consistent with the topology preserving property of the PDE.

The key to the method of Han et al. is the definition of a *simple point*. A point is said to be *simple* if its addition to or removal from a digital object does not change the object topology. A local condition to determine if a point is simple is proved in [4]. A point x is simple iff the number of connected components of the foreground in a 3×3 neighborhood of x is equal to the number of connected components in the background, and both numbers are one. Note that when defining connectivity, the foreground and background connectivity must be a *connectivity pair*, that is, a 4 pixel

neighborhood defined for connectivity of the foreground and an 8 pixel neighborhood for the background, or vice-versa. The algorithm of Han et al. updates each grid point of the level set function successively and checks each grid point to see whether the point is simple; if the point is simple then the level set function at the point is updated ordinarily; if not and the point would change sign, then the point is updated as far as possible without changing sign, thus preventing a discrete topology change.

It should be noted that the method of Han et al. can also be used in conjunction with the polygon evolution or marker particle method described in Section 2.3.1. To do this, one simply uses an artificial grid that is the size of the image to be segmented; the grid indicates whether or not the grid point is inside or outside the contour or polygon. Note that this array can be updated efficiently by noting that only points in a one pixel neighborhood of the polygon can change values if the time step is chosen (as is usually done).

2.3.3 Alternative Implementation using FFT

In this section, we describe implementing the gradient descent of an energy of the form

$$E(C) = \frac{1}{2} \iint_{C \times C} \phi(C(s), C(\hat{s}))\, \mathrm{d}\hat{s}\, \mathrm{d}s$$

where $\phi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is symmetric and is a function of the Euclidean distance $\|C(s) - C(\hat{s})\|$. The gradient descent flow of such an energy is $C_t = -(\vec{e} \cdot \mathcal{N})\mathcal{N} + \kappa p \mathcal{N}$ where

$$\vec{e}(s) = \int_C \nabla_1 \phi(C(s), C(\hat{s}))\, \mathrm{d}\hat{s} \quad \text{and} \quad p(s) = \int_C \phi(C(s), C(\hat{s}))\, \mathrm{d}\hat{s} \tag{49}$$

We give an efficient scheme to implement the level set version of this evolution, which does not require the polygon extraction steps described in Section 2.3.2.1. This method may be of use when ϕ is chosen to be non-singular. For example, if we choose $\phi(\mathbf{x}, \mathbf{y}) = 1/\sqrt{\|\mathbf{x} - \mathbf{y}\|^2 + \epsilon^2}$ where $\epsilon > 0$, we get a non-singular "approximation" to

the singular function $1/\|\mathbf{x}-\mathbf{y}\|$. In this case, we find that

$$\begin{aligned} \vec{e}(s) &= \int_C \frac{C(s)-C(\hat{s})}{(\|C(s)-C(\hat{s})\|^2+\epsilon^2)^{3/2}}\ \mathrm{d}\hat{s} \\ p(s) &= \int_C \frac{\mathrm{d}\hat{s}}{\sqrt{\|C(s)-C(\hat{s})\|^2+\epsilon^2}}. \end{aligned} \tag{50}$$

These terms may be considered approximations to $\vec{E}$ and P (with the log term ignored) (47), respectively. Note that the gradient no longer becomes infinite as a curve approaches self-intersection, and therefore it cannot be guaranteed that topology is preserved in the presence of an image-based term. However, we observed experimentally that this flow has similar properties to (36) and may preserve topology. Again, this method is probably more useful for applications where a non-singular ϕ is needed, not necessarily for topology preservation.

The functions p and $\vec{e}$, which are integrals around C of some quantity, may be written as region integrals over the image domain $\Omega \subset \mathbb{R}^2$ with the use of the Dirac distribution. Define $\mathscr{P} : \Omega \to \mathbb{R}$ as

$$\mathscr{P}(x) = \int_\Omega (\delta \circ \Psi)(y) K(x-y) \|\nabla \Psi(y)\|\ \mathrm{d}y \tag{51}$$

where δ is the one dimensional Dirac distribution, Ψ is the level set function of the curve C, $K : \Omega \to \mathbb{R}$ is

$$K(\xi) = \phi(\xi, 0) (= \frac{1}{\sqrt{\|\xi\|^2+\epsilon^2}} \text{ in the special case}),$$

and $\mathrm{d}y$ denotes Lebesgue measure on $\mathbb{R}^2$. With this definition, we can see $\mathscr{P}|C = p$ where $\mathscr{P}|C$ denotes the restriction of $\mathscr{P}$ to C. A similar expression to (51) can be obtained for both components of $\vec{e}$. Note that (51) is in the form of a convolution, and therefore it follows that

$$\mathscr{P} = (\delta \circ \widehat{\Psi \|\nabla \Psi\|} \widehat{K})^\vee \tag{52}$$

where $\widehat{\cdot}$ denotes the Fourier transform of $\cdot$, and $\cdot^\vee$ denotes the inverse Fourier transform of $\cdot$.

Since we are implementing the curve evolution on a finite grid, we must use a discrete equivalent to the continuous equations given above. If we assume that the grid spacing is one in both the x and y directions, then (51) remains valid if we change the measure $\mathrm{d}y$ to two dimensional counting measure. Equation (52) remains valid, however, we note that the Fourier integrals are with respect to counting measure. To represent $\delta \circ \Psi$ on a finite grid, we use an approximation of δ. The approximation is simply a function that is peaked at zero, and integrates to one over $\mathbb{R}$ (see [17] for some examples). Fast computation of the discrete Fourier and inverse Fourier transforms are done using the FFT. We are now ready to describe the algorithm for computing P and $\vec{E}$ on the narrow band of the level set function Ψ.

1. Compute $\widehat{K}$ using the FFT, and store it. This will only need to be done once since K does not depend on the curve.

2. Compute $\delta \circ \Psi \|\nabla \Psi\|$, and then $\delta \circ \widehat{\Psi \|\nabla \Psi\|}$ using the FFT.

3. Multiply $\widehat{K}$ with $\delta \circ \widehat{\Psi \|\nabla \Psi\|}$ and compute the inverse Fourier transform using the FFT. This gives $\mathscr{P}$ defined on the whole grid.

4. Interpolate $\mathscr{P}$ to find $\mathscr{P}$ on the zero level set of Ψ, which is actually p.

5. Extend p to the narrow band of Ψ.

The computation of $\vec{e}$ follows the same steps as for p, but K is replaced with the function, $\tilde{K} : \Omega \to \mathbb{R}^2$,

$$\tilde{K}(\xi) = \nabla_1 \phi(\xi, 0) = \left(\frac{\xi}{(\|\xi\|^2 + \epsilon^2)^{3/2}}\right).$$

This algorithm is definitely easier to implement than steps 1 and 2 of the algorithm in the previous section. Efficiently finding a polygonal estimate of the zero level set and organizing it in such a way that p and $\vec{e}$ can be easily computed is a little laborious. Although we do not consider the case of evolving surfaces in this work, we can say

finding a polyhedral estimate of the zero level set is much more of a programming hassle than the case of curves. In this case, an algorithm similar to marching cubes [61] is needed. The runtime of steps 1 and 2 of the algorithm in the previous section is $O(N^2)$ where N is the number of vertices of the polygonal estimate of the zero level set. The number of vertices is dependent on the length of the curve, and the grid size. The runtime of the algorithm presented in this section is $O(N^2 \log N)$ where in this case N^2 is the grid size of the image. Thus, if the length of the curve is small relative to the size of the grid then the algorithm of the previous section is faster; if the length of the curve is large relative to the size of the grid, then the algorithm presented in this section is faster.

2.4 Simulations

In this section, we present some results illustrating the geometric properties of the proposed topology preserving flows, and we also present examples illustrating the use of the topology preserving flows in segmentations of real images.

2.4.1 Geometric Properties

The top row of Fig. 3 shows the evolution of a polygonal spiral evolving solely under the polygon topology preserving flow (25). Notice that the spiral unravels from the inside by shrinking its inner segments and pushing segments into a small area. Although it looks as if several vertexes of the polygon have collapsed to a single point in the evolution, this cannot happen as was proved in Section 2.2.1.3. The polygon then becomes convex, and finally converges to a regular polygon. We should note that the polygon does not shrink to a point, but converges to a regular polygon of some fixed perimeter. This is expected because of the term, $\sum_{i\in\mathbb{Z}_n}(|C_i|\ln|C_i| - |C_i|)$. Based on several other experiments, which are not shown here in the interest of space, we believe that any polygon converges to a regular polygon under this flow, but this has not been proven. Therefore, this flow can also be used as a global regularity term in

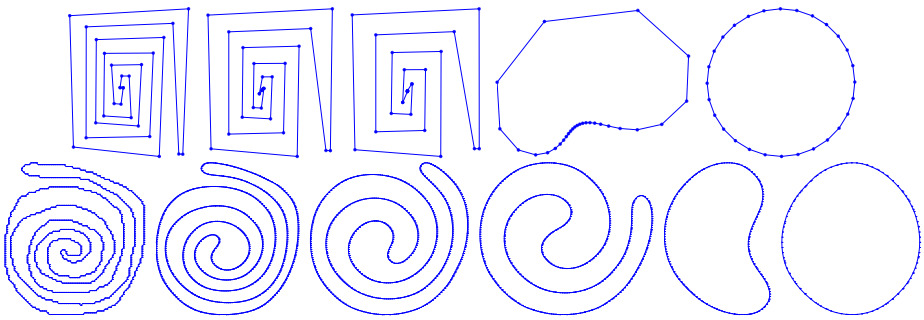


Figure 3: Snapshots of an evolution of a polygonal spiral (top) under (25), and evolution of a spiral (bottom) under (45). The polygon is evolving solely under the proposed topology preserving flow. For visibility, the spatial scale is shrunk as the evolution progresses.

addition to preserving topology of active polygons.

The bottom row of Fig. 3 shows the evolution of a spiral evolving solely under the topology preserving flow (45). The spiral eventually becomes convex, becomes circular, and shrinks to a configuration that is not representable with the given resolution of the level set function. The behavior of this flow, as witnessed through this experiment, shows some similarities to curve shortening flow. In particular, the topology preserving flow has smoothness properties like the curve shortening flow. It also unwinds the spiral without crossing itself like the curve shortening flow. Note that it is shown in [1] that the energy E_c is minimized by a circle among curves having a constant perimeter, which is similar to the energy that is defined as the length of a closed curve. Fig. 4 shows an obvious difference between (45), which is shown on the bottom, and the curve shortening flow, which is shown on the top. In this figure, the evolution of a thin, long curve is shown. Notice, in the bottom of Fig. 4, that the curve becomes thicker as parallel sides are pushed apart. This is due to the term $\mathcal{E}_\epsilon \mathcal{N}$ in the curve evolution of (45). Since these parallel sides have nearly zero curvature, the curve shortening flow cannot not push the sides apart. For topology preservation, this repulsion is a desirable property.

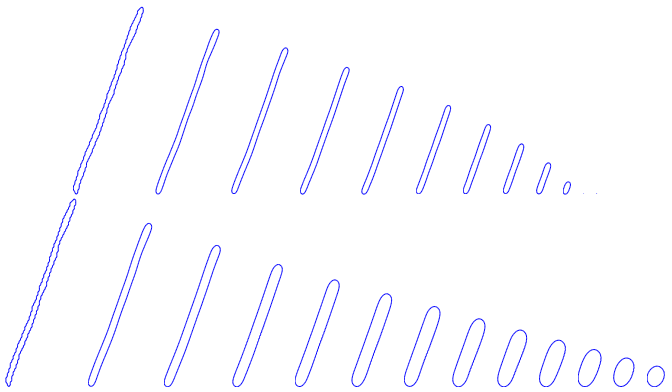


Figure 4: Illustration of a difference between curvature flow (top) and proposed flow (45) (bottom).

2.4.2 Image Segmentations

In all the following image segmentations, we have used the flow considered in [17] as the image-based term. The flow, in summary, moves a contour or polygon to separate an image into two regions that are piecewise constant.

In all the following results, it should be noted that any weight on the proposed flow can be used to preserve topology. However, the lower the weight on the topology preserving term, the greater the possibility for dependence on the method of Han et al. The weights are given so that the results can be reproduced. Results of segmentation of a simple leaf image with an active polygon are shown in Fig. 5. The result of running the Chan and Vese [17] flow without the proposed topology preserving flow is shown on the top row of Fig. 5. Notice that a topology change occurs. From the nature of the Chan-Vese flow, the vertices indicated by the arrows in Fig. 5 will move the polygon toward self intersection. This is because the vertices will move in the direction that reduces the white area as fast as possible; clearly, the direction of the vertices that decreases white area fastest moves the polygon toward self intersection. After self intersection, the normal vectors become flipped at the indicated vertices, and the flow moves in the wrong direction. Since there is no way to consistently define a normal vector field when the polygon is no longer embedded, this problem cannot

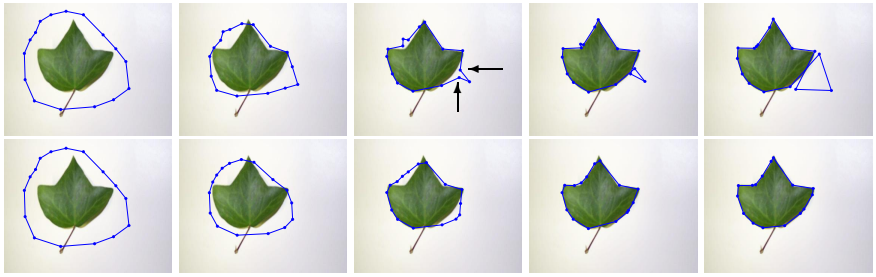


Figure 5: Snapshots of an evolution of active polygon using Chan-Vese flow with the proposed topology preserving flow.

be corrected. The bottom row of Fig. 5 shows snapshots of the evolution of Chan-Vese flow weighted 98% and the topology preserving flow weighted 2%. Notice the continuous motion of the polygon to keep the polygon away from self intersection. The polygon is more "regular" during the evolution than the polygon that results without using the proposed flow, and this verifies that the flow serves to globally regularize the polygon.

The next simulations are using active contours for image segmentation. Figure 6 shows the results of segmentation of an image with two closely spaced bones with no topology preservation, topology preservation of Han et al. [44] alone, and the proposed method. The top row is the result with a curvature regularization term; the curvature term is weighted 50% to compensate for the noise. As can be seen, the two contours merge across the thin gap between the two bones. This merging is due to the nature of the image and Chan-Vese flow, and not a step size problem. The middle row shows the result of the topology preserving method of Han et al. with curvature weighted 50%. The final row shows the result with our proposed flow weighted 27% and the remaining percent is the image-based term. It should be noted that 27% is not a "magic number", and that a wide range of weights give similar segmentations. Moreover, any weight theoretically prevents topology change without the use of the method of Han et al. A zoom of the region separating the two bones is shown in Fig. 7. Notice that the contour is stopped abruptly in an arbitrary location

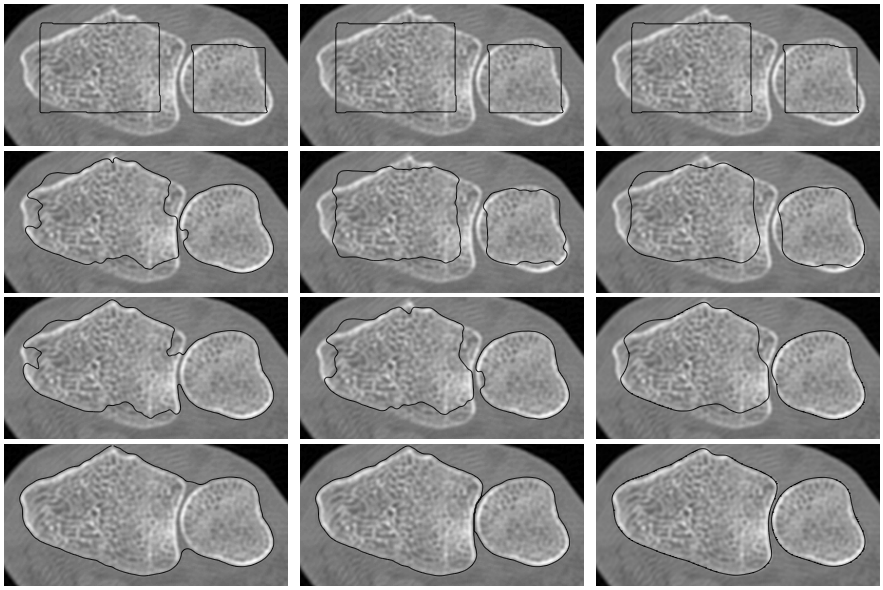


Figure 6: Segmentation of bones CT image with no topology preservation (left), with topology preservation of Han et al. alone (middle), and with proposed topology preservation (right).

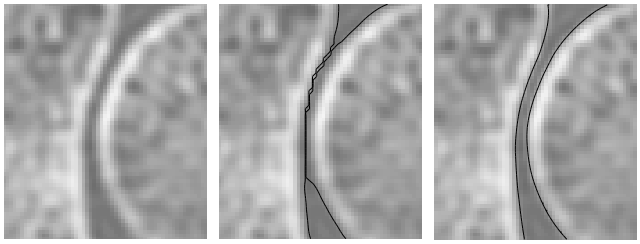


Figure 7: Gap between bones with contours overlayed. Results of no topology preservation, topology preservation of Han alone, and proposed topology preservation.

relative to the image features in the result of the method of Han et al. without the proposed flow. These kind of artifacts are typical of a forced topology constraint that does not model the actual behavior of the underlying PDE. In contrast, using the proposed flow globally regularizes the evolving contour and thereby pushes the contour to attract relevant image features. Also, since the proposed flow gradually adjusts the curve motion away from topology change, the curve is never one pixel away from self-intersecting, and thus the discrete topology preservation step never needs to be invoked in this example.

Figure 8 shows the evolution of an active contour segmenting a simple image of

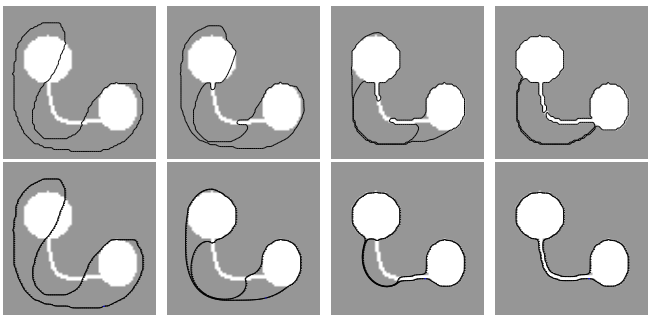


Figure 8: Geometrical pathology typical of forced topology preservation constraint of Han et al. alone (top), and the correct segmentation by the proposed geometric regularization flow.

two circles connected by a thin strip. The top row shows the result of the topology preserving level set method of Han et al. alone, and the bottom row shows the results of the proposed method. Note that we have used curvature regularization (weighted 10%) in the method of Han to keep the curve smooth. Although the topology is correct, the segmentation is geometrically inaccurate. One may argue that weighting the curvature high enough in the method of Han et al. alone will correct this problem. However, weighting the curvature just high enough to overcome the pathology, produces a flow that looks like curvature flow, and fails to even capture the circles in the image. The bottom row of Fig. 8 shows the result of segmentation with Chan-Vese flow weighed 96% and the proposed flow weighted 4%. The curve repels from itself in a continuous manner as it becomes close to self intersection. The curve is then pushed in the right direction to capture the thin strip as the thin strip slowly becomes detected by the image-based term. This experiment serves to illustrate the geometrical pathologies that generally occur when a discrete topology constraint is artificially forced on a PDE that exhibits no topology preserving properties. The experiment also illustrates that a true topology preserving PDE with global geometric regularization reduces the possibilities of these geometric pathologies.

A simulation illustrating the effect of the weighting factor of the proposed flow on the final segmentation is shown in Fig. 9. The initial contour is shown in (a).

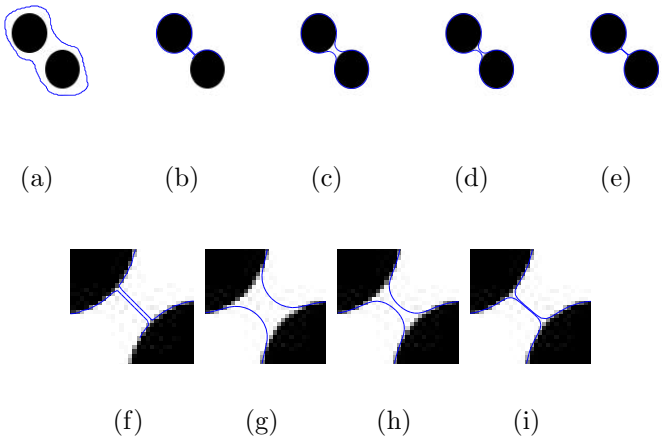


Figure 9: The effect of changing the weight on the topology preserving term. (a) Initial contour, (b) and (f) result of Han *et al.*, (c) and (g) 33%, (d) and (h) 20%, (e) and (i) 7% weights on proposed topology preserving term. (f), (g), (h). (i) are enlargements of the middle area between the circles.

The result of Han et al. algorithm with Chan-Vese flow weighted 70% and curvature regularization weighted 30% is shown in (b), and the region between the two circles is enlarged in (f). Final results of the proposed flow weighted 33%, 20% and 7% along with Chan-Vese flow are given in (c)-(e) and (g)-(i), respectively. In the simulations showing the proposed flow, the portion of the contour between the circles stops when the Chan-Vese force balances the proposed force. Weighting the proposed term lower gives a result similar to using Han *et al.*'s method alone. Increasing the weight of the proposed term ensures the final segmentation is increasingly further from changing topology, and is more globally regular. It should be noted that the proposed flow affects the original image-based evolution significantly when the curve is close to self intersection, and is highly irregular.

2.5 Conclusion

In summary, we have presented a novel method for introducing global regularity into contour and polygon evolutions and simultaneously enforcing a prior assumption on

the topology of objects to be detected from an image. This was done by constructing a geometric flow that can be added to existing image-based evolutions of active polygons or contours. We have demonstrated that the contructed flow through its global regularization and gradual topology preservation reduces the possibility of geometric pathologies that plague discrete topology constraints forced on PDEs having no tendency to preserve topology. In particular, we noted that the justified way for using a discrete topology constraint is in conjunction with a PDE that preserves topology as an appropriate numerical technique (much as conservation laws are enforced through proper numerical techniques). We have given analytical justification that the constructed geometric flow preserves topology in the presence of arbitrary bounded image-based forces in continuous time. Experiments verified the topology preserving properties of the proposed flow, and showed that global regularization and gradually adjusting topology produces geometrically accurate segmentations that are often needed in applications.

CHAPTER III

SOBOLEV ACTIVE CONTOURS

All previous geometric active contour models that have been formulated as gradient flows of various energies use the same L^2-type inner product to define the notion of gradient. Recent work has shown that this inner product induces a pathological Riemannian metric on the space of smooth curves. However, there are also undesirable features associated with the gradient flows that this inner product induces. In this chapter, we reformulate the generic geometric active contour model by redefining the notion of gradient in accordance with Sobolev-type inner products. We call the resulting flows Sobolev active contours. Sobolev metrics induce favorable regularity properties in their gradient flows. In addition, Sobolev active contours favor global translations, but are not restricted to such motions; they are also less susceptible to certain types of local minima in contrast to traditional active contours. These properties are particularly useful in tracking applications. We demonstrate the general methodology by reformulating some standard edge-based and region-based active contour models as Sobolev active contours and show the substantial improvements gained in segmentation.

3.1 Introduction

Throughout the history of active contours, there has been much interest in defining new (and more complicated) energies to drive these contours. The need for more complicated energies is not only because of the need to segment increasingly more complicated images, but also in part due to the optimization technique used to optimize these energies, which often times becomes trapped in *local* minima. With few exceptions, all active contour methods use the same optimization technique based on

gradient descent. However, as noted in Chapter 1, [68, 116] observed that all previous works on geometric active contours that derive *gradient flows* to minimize energies imply a natural notion of distance induced by a Riemannian structure. Indeed, the notion of gradient of such energies relies on an inner product defined the set of perturbations of a curve. All of these previous works on geometric active contours use the same geometric L^2-type inner product, which we refer to as H^0, to define gradient. However, [68, 113] have shown a surprising property: the Riemannian metric on the space of curves induced by the H^0 inner product is pathological, i.e., the "distance" between *any* two curves is zero.

In addition to the pathologies of the Riemannian structure induced by H^0, there are also undesirable features of H^0 gradient flows as we will demonstrate in this chapter. Some of these features are listed below.

1. There are no regularity terms in the definition of the H^0 inner product. That is, there is nothing in the definition of H^0 that discourages flows that are not smooth in the space of curves. By smooth in the spaces of curves, we mean that the surface formed by plotting the evolving curve as a function of time is smooth. Thus, when energies are designed to depend on the image that is to be segmented, the H^0 gradient is very sensitive to noise in the image. As a result, the curve becomes non-smooth instantly. Therefore, in geometric active contours models, a penalty on the curve's length is added to keep the curve smooth in addition to producing a variational problem that is well-posed. However, this changes the energy that is being optimized. Moreover, as we shall demonstrate, the length penalty is sometimes insufficient, and leads to other problems.

2. H^0 gradients, evaluated at a particular point on the curve, depend locally on derivatives of the curve. Therefore, as the curve becomes non-smooth, as mentioned above, the derivative estimates become inaccurate, and thus, the curve

evolution becomes inaccurate. Moreover, for region-based and edge-based active contours, the H^0 gradient at a particular point on the curve depends locally on image data at the particular point. Although region-based energies may depend on global statistics, such as mean values of regions, the H^0 gradient still depends on local image data. These facts imply that the H^0 is sensitive to noise and local features.

3. The H^0 norm gives *non-preferential* treatment to arbitrary deformations regardless of whether the deformations are global motions (not changing the shape of the curve) such as translations, rotations and scales or whether they are more local deformations. This implies, as we shall show, that the H^0-gradient of image dependent energies "encourages" points on the evolving curve to move "independently" to decrease energy rather than encouraging the points to move more collectively. This restricts the gradient at a particular point from "seeing" information located at other points of the curve. Therefore, the curve evolving according to the H^0 gradient flow is susceptible to local minima of the energy.

4. Many geometric active contours (such as edge and region-based active contours) require that the unit normal to the evolving curve be defined. As such, the evolution does not make sense for polygons. Moreover, since in general, an H^0 active contour does not remain smooth, one needs special numerical schemes based on viscosity theory in a level set framework to define the flow.

5. If the energy depends on n derivatives of the curve, then the H^0 gradient has $2n$ derivatives of the curve. Since the corresponding level set flows with higher than two derivatives are not known to have a maximum principle, level set methods [78] are difficult to use for the numerical implementation [23]. Therefore, one may try to use particle methods to implement the flow. However, flows with higher than two derivatives are generally difficult to implement because of

numerical artifacts.

6. Lastly, as a specific example, the H^0 gradient ascent for arclength, i.e., backward heat flow, is ill-posed. This is quite odd in an intuitive manner because there is nothing in the definition of length itself that indicates that a flow to increase length should be ill-posed. We shall see that this ill-posedness is a property of H^0 not just the energy itself.

In this chapter, we consider using inner products arising from Sobolev spaces to define gradients. Unlike the research done in active contours for the past 20 years, which has focused on finding various *new energies* to deal with local minima and other problems, we present a *new way* of doing active contours by giving an alternative way of minimizing active contour energies.

By changing the Riemannian metric associated with the space of curves, we are able to regularize the minimizing flows associated with active contour energies without requiring the addition of regularization penalties in the aforementioned energies. Indeed, the change of metric does not affect the global minima of the energy, but it changes the notion of gradient, and the notion of "neighborhood of a curve". While previous local minimizers continue to be local minimizers (or at least critical points) in the Sobolev metrics, the definition of locality is completely different. As a result, Sobolev active contours are much more robust to the local minima which strongly influence previous, standard, active contours. Moreover, any existing active contour method can get the added benefits of this new approach with minimal changes to its existing implementation and with little extra computational time.

We would like to point out that Sobolev gradient methods have been used in the past; for example the book [74] (see also references within) presents the Sobolev gradient and applies it to the numerical solution of various physical problems. More recently, [12] has examined using various Sobolev gradients (especially fractional

Sobolev gradients) to various inverse problems of boundary reconstruction, and examined rates of convergence among the various gradients considered. We and (independently) [19] have introduced the Sobolev method to active contour problems. Charpiat et al. [19] also go on to consider other "coherent" motions resulting from various inner products (see also the work of [64] for a related idea).

3.2 General Theory

In the next sections, we begin by using principles from Riemannian geometry [30, 57] and show how they fit into the framework of geometric active contours, which is essential to construct the Sobolev active contour.

3.2.1 Structure on the Space of Curves

Let M denote the set of smooth immersed curves in $\mathbb{R}^d$, where $d \geq 2$. We show how to make M a differentiable manifold. Let us give a precise definition of M:

$$M := \{c \in C^\infty(S^1, \mathbb{R}^d) : c'(\theta) \neq 0 \ \forall \, \theta \in S^1\},$$

where S^1 denotes the circle. Denote by F the Fréchet space $C^\infty(S^1, \mathbb{R}^d)$ of smooth vector fields on S^1 whose topology is defined by the following seminorms:

$$p_n(h) = \sup_{\theta \in S^1} |h^{(n)}(\theta)|, \quad n = 0, 1, \ldots$$

where $h \in F$. We now show

Proposition 3.2.1. *M is a differentiable manifold whose model space is the infinite dimensional space, F.*

Proof. 1. We first define charts from M to F. Let $c \in M$, then denote by $\Phi_c^{-1} : \tilde{U}_c \subset F \to U_c \subset M$ a chart at $c \in M$ defined by $\Phi_c^{-1}(h) = c + h$. We must define the neighborhood $\tilde{U}_c$, and show that $c + h \in M$ where $h \in \tilde{U}_c$. Choose

$$0 < \epsilon < \inf_{\theta \in S^1} |c'(\theta)|.$$

Note the choice of ϵ is possible since $|c'(\theta)| \neq 0$ ($c \in M$), and the infimum cannot be zero since $|c'|$ is continuous on the compact set S^1. Now define $\tilde{U}_c := \{h \in F : p_1(h) < \epsilon\}$, which by definition of F is open. It is now clear that $c + h \in M$ for $h \in \tilde{U}_c$. It is also clear that Φ_c is a bijection.

2. We now show that the charts defined above are compatible. Let $c, \tilde{c} \in M$, $c \neq \tilde{c}$ and $U_c \cap U_{\tilde{c}} \neq \emptyset$. First note that $\Phi_c(U_c \cap U_{\tilde{c}}) = \{h \in F : p_1(h) < \epsilon_c, p_1(h - (\tilde{c} - c)) < \epsilon_{\tilde{c}}\}$, which is clearly open. Next, we see that $\Phi_{\tilde{c}} \circ \Phi_c^{-1} : \Phi_c(U_c \cap U_{\tilde{c}}) \to \Phi_{\tilde{c}}(U_c \cap U_{\tilde{c}})$ is given by

$$(\Phi_{\tilde{c}} \circ \Phi_c^{-1})(h) = (c - \tilde{c}) + h,$$

which is clearly a C^∞ diffeomorphism.

□

Remark 3.2.1. *Note, we may consider* $M = \{c \in C^1(S^1, \mathbb{R}^d) : c'(\theta) \neq 0 \, \forall \theta \in S^1\}$ *and* F *to be the Banach space* $C^1(S^1, \mathbb{R}^d)$ *with topology defined by the norm* $\|h\| = \sup_{\theta \in S^1}(|h(\theta)| + |h'(\theta)|)$. *In this case,* M *is also a differentiable manifold. However, the tangent to the curve,* c' *is no longer in the model space. See [67] for more details.*

Remark 3.2.2. *Let* $\mathrm{Diff}(S^1)$ *be the set of smooth automorphisms of* S^1. *Note that in our definition of the manifold* M, *the space of* immersed *curves, two different parameterizations of a curve (i.e.,* $c_1, c_2 \in M$ *such that* $c_1 = c_2 \circ \phi$ *where* $\phi \neq id_{S^1}$ *is in* $\mathrm{Diff}(S^1)$*) are considered two distinct elements of* M. *For computer vision applications one should consider the quotient space*

$$B := M/\mathrm{Diff}(S^1);$$

this quotient models the space of geometrical curves, *which are "curves up to a choice of parametrization". This space, however, is not a manifold [68]; but a slightly smaller quotient space*

$$B_f := \mathrm{Imm}_f/\mathrm{Diff}(S^1)$$

of the freely immersed curves is a manifold. In our work, we will use M. The results we obtain, however, are geometric *in that they do not depend on a particular parameterization of a curve; so they may be "projected" to B_f.*

For $c \in M$, we denote by T_cM the tangent space of M at c. Since M is an open subspace of $C^\infty(S^1, \mathbb{R}^d)$, the tangent space may be identified with $C^\infty(S^1, \mathbb{R}^d)$ itself. It is easy to see that this identification satisfies all standard requirements and/or different choices of definitions for T_cM.

3.2.2 Inner Products on T_cM

We now define inner products on T_cM.

Definition 3.2.1. *Let $c \in M$, L be the length of c, and $h, k \in T_cM$. Let $\lambda > 0$, and $n \in \mathbb{N}$. We assume h and k are parameterized by the arclength parameter of c.*

1. $\langle h, k \rangle_{H^0} := \dfrac{1}{L} \displaystyle\int_c h(s) \cdot k(s) \, \mathrm{d}s$
2. $\langle h, k \rangle_{H^n} := \langle h, k \rangle_{H^0} + \lambda L^{2n} \left\langle h^{(n)}, k^{(n)} \right\rangle_{H^0}$
3. $\langle h, k \rangle_{\tilde{H}^n} := avg(h) \cdot avg(k) + \lambda L^{2n} \left\langle h^{(n)}, k^{(n)} \right\rangle_{H^0}$

where $\int \cdot \, \mathrm{d}s$ is the integral w.r.t the arclength parameter, $avg(h) := \dfrac{1}{L} \displaystyle\int_c h(s) \, \mathrm{d}s$, and $h^{(n)}$ denotes the n^{th} derivative of h with respect to arclength.

It is easy to verify that the above definitions are inner products. Note that we have introduced length dependent scale factors so that the above inner products (and the corresponding norms) are independent of curve rescaling (we do not want a rescaling of the curve to change the inner product).

Remark 3.2.3. *The most general scale invariant definition of H^n is*

$$\langle h, k \rangle_{H^n} = \langle h, k \rangle_{H^0} + \sum_{j=1}^{n} \lambda_j L^{2j} \left\langle h^{(j)}, k^{(j)} \right\rangle_{H^0} \tag{53}$$

where $\lambda_j \geq 0$ for $j = 1, \ldots, n-1$, and $\lambda_n > 0$. We will see in Section 3.2.4 that it is unnecessary for our purposes to use this general definition.

Remark 3.2.4. *Other definitions of H^n and $\tilde{H}^n$ are possible [113]. We illustrate the ideas on planar curves ($d = 2$). We recall from Remark 3.2.2 that one main goal of the present theory is to be* geometric*; that is, any result may be projected to the space B_f of* geometrical curves*. A possible choice for representing the tangent T_cB_f is $T_cM/\sim$ where $h \sim k$ iff $h \cdot \mathcal{N} = k \cdot \mathcal{N}$ where $\mathcal{N}$ is the inward normal of c. The other definitions are as follows: Let $\alpha, \beta : S^1 \to \mathbb{R}$ be the normal components of two perturbations h, k (i.e., $\alpha = h \cdot \mathcal{N}$ and $\beta = k \cdot \mathcal{N}$) then*

$$\langle [h], [k] \rangle_{H^{1^*}} := \frac{1}{L} \int_c \left[\alpha(s)\beta(s) + \lambda L^2 \alpha'(s)\beta'(s) \right] \, \mathrm{d}s \tag{54}$$

$$\langle [h], [k] \rangle_{H^{1^+}} := \frac{1}{L} \int_c \left[\alpha(s)\beta(s) + \lambda L^2 (\alpha(s)\mathcal{N}(s))' \cdot (\beta(s)\mathcal{N}(s))' \right] \, \mathrm{d}s. \tag{55}$$

The inner product in (55) *can be simplified as follows*

$$\langle [h], [k] \rangle_{H^{1^+}} = \frac{1}{L} \int_c \left[(1 + \lambda L^2 \kappa^2(s))\alpha(s)\beta(s) + \lambda L^2 \alpha'(s)\beta'(s) \right] \, \mathrm{d}s. \tag{56}$$

Ignoring the $\alpha'\beta'$ term and the length factors, (56) *becomes the same inner product considered in [68] for shape analysis.*

One of the important consequences of defining an inner product on T_cM is that, with some other smoothness and compatibility conditions, it produces a *Riemannian* structure on M. This means that specifying an inner product on T_cM induces a distance between points on M. The authors in [68, 113] consider the Riemannian structure induced from the H^0 inner product and show that it is pathological; they also propose alternative inner products to achieve a non-trivial Riemannian distance. We explore this idea in [67] (see also [117]).

3.2.3 Gradient of Functionals on M

We now define the notion of gradient of a functional $E : M \to \mathbb{R}$.

Definition 3.2.2. *Let $E : M \to \mathbb{R}$.*

1. *If $c \in M$ and $h \in T_cM$, then the* variation of E *in the direction h is*

$$\mathrm{d}E(c) \cdot h = \frac{\mathrm{d}}{\mathrm{d}t} E(c + th)\bigg|_{t=0},$$

where $(c + th)(\theta) := c(\theta) + th(\theta)$ and $\theta \in S^1$. Note $\mathrm{d}E(c) \in T_c^*M$ is a linear operator on T_cM, $\mathrm{d}E(c)$ is called the differential at c.

2. *Assume $\langle , \rangle_c$ is an inner product on T_cM. The* gradient of E *is a vector field $\nabla E(c) \in T_cM$ that satisfies*

$$\mathrm{d}E(c) \cdot h = \langle h, \nabla E(c) \rangle_c$$

for all $h \in T_cM$.

We now give an intuitive interpretation of the gradient, which tells us the significance of this perturbation. We show that the gradient is the most efficient perturbation; that is, the gradient maximizes the change in energy per "cost" of perturbing the curve.

Proposition 3.2.2. *Let $\| \cdot \|_c$ be the norm induced from the inner product $\langle \cdot, \cdot \rangle_c$ on T_cM. Suppose $\mathrm{d}E(c) \neq 0$, and the gradient with respect to $\langle \cdot, \cdot \rangle_c$ exists; then the problem*

$$\sup_{\{h \in T_cM, \|h\|_c = 1\}} \mathrm{d}E(c) \cdot h = \sup_{\{k \in T_cM, k \neq 0\}} \frac{\mathrm{d}E(c) \cdot k}{\|k\|_c} \tag{57}$$

has a unique solution, $k = \nabla E(c) \in T_cM, h = k/\|k\|$.

Proof. Note $\mathrm{d}E(c) \cdot h = \langle \nabla E(c), h \rangle_c \leq \|\nabla E(c)\|_c \|h\|_c$ by the Cauchy-Schwartz inequality; moreover, $\mathrm{d}E(c) \cdot (\nabla E(c)) = \|\nabla E(c)\|_c^2$. □

If the gradient of E exists, then by the proposition above, we have that $k = \nabla E(c)$ attains the supremum on the right hand side of (57). Note for $\lambda \to +\infty$, translations have negligible norm with respect to other directions in the tangent space, that is, if h is a translation (i.e. it is constant w.r.t. s) and k is not a translation, then

$\lim_{\lambda\to+\infty} \|h\|/\|k\| = 0$, both when $\|\cdot\|$ is the norm induced by H^n and when it comes from $\tilde{H}^n$. In light of the interpretation of the gradient as the perturbation that attains the supremum in (57), it follows that translations are favored for gradients in H^n and $\tilde{H}^n$ when λ is large (if they reduce energy).

Remark 3.2.5 (Comment on H^n for $n \geq 2$). *Translations are favored for H^n and $\tilde{H}^n$ gradients when $\lambda \to +\infty$. This can be quite important for tracking applications where the object to be tracked is usually translating. One may wonder whether using higher order Sobolev inner products, H^n and $\tilde{H}^n$ for $n \geq 2$, will favor higher order polynomial motions of degree n. Note however, that any polynomial perturbation defined on S^1, the circle, must be constant to conform to periodic boundary conditions. Thus, higher than order $n = 1$ Sobolev gradients also favor just translations. In this sense, there is not an advantage, with respect to gaining higher dimensional preferred motions, in using higher order Sobolev gradients. However, one gains added regularity of the gradient flow in using higher order Sobolev gradients.*

Remark 3.2.6. *For the inner product on $T_cM/\sim$ defined in* (54)*, it follows that dilations (i.e., perturbations of the form $h = \pm\mathcal{N}$) are favored for gradients when λ is large.*

3.2.4 Relation Between H^n and $\tilde{H}^n$

We show that the norms associated with the inner products H^n and $\tilde{H}^n$, *i.e.*,

$$\|h\|_{H^n} = \sqrt{\int_0^L \frac{1}{L}|h(s)|^2 + \lambda L^{2n-1}|h^{(n)}(s)|^2 \,\mathrm{d}s} \tag{58}$$

$$\|h\|_{\tilde{H}^n} = \sqrt{|\mathrm{avg}(h)|^2 + \lambda L^{2n-1}\int_0^L |h^{(n)}(s)|^2 \,\mathrm{d}s} \tag{59}$$

are equivalent.

Proposition 3.2.3 (Equivalence of H^n and $\tilde{H}^n$). *The norms defined by* (58) *and* (59) *on T_cM are topologically equivalent; that is,*

$$\alpha\|h\|_{\tilde{H}^n} \leq \|h\|_{H^n} \leq \beta\|h\|_{\tilde{H}^n}$$

for all $h \in T_cM$, and $\alpha, \beta > 0$ are not a function of $c \in M$ and $h \in T_cM$.

Proof. We first derive a simple Poincaré inequality. For convenience, we will replace $h : S^1 \to \mathbb{R}^d$ with its periodic extension $h : \mathbb{R} \to \mathbb{R}^d$. We have

$$h(u) - h(0) = \int_0^u h'(s)\,\mathrm{d}s = -\int_u^L h'(s)\,\mathrm{d}s,$$

since h is periodic. Now,

$$\begin{aligned}
h(u) - h(0) &= \frac{1}{2}\left(\int_0^u h'(s)\,\mathrm{d}s - \int_u^L h'(s)\,\mathrm{d}s\right) \Rightarrow \\
\mathrm{avg}(h) - h(0) &= \frac{1}{2L}\int_0^L \left(\int_0^u h'(s)\,\mathrm{d}s - \int_u^L h'(s)\,\mathrm{d}s\right)\mathrm{d}u \Rightarrow \\
|\mathrm{avg}(h) - h(0)| &\leq \frac{1}{2L}\int_0^L \left(\int_0^u |h'(s)|\,\mathrm{d}s + \int_u^L |h'(s)|\,\mathrm{d}s\right)\mathrm{d}u \\
&= \frac{1}{2L}\int_0^L \left(\int_0^L |h'(s)|\,\mathrm{d}s\right)\mathrm{d}u = \frac{1}{2}\int_0^L |h'(s)|\,\mathrm{d}s
\end{aligned}$$

so that by replacing 0 with an arbitrary point, we have

$$\sup_u |h(u) - \mathrm{avg}(h)| \leq \frac{1}{2}\int_0^L |h'(s)|\,\mathrm{d}s. \tag{60}$$

Now, using (60) and Hölder's inequality, we see

$$\begin{aligned}
\sqrt{\int_0^L |h(s) - \mathrm{avg}(h)|^2\,\mathrm{d}s} &\leq \sqrt{L}\sup_u |h(u) - \mathrm{avg}(h)| \\
&\leq \sqrt{L}\int_0^L |h'(s)|\,\mathrm{d}s \\
&\leq L\sqrt{\int_0^L |h'(s)|^2\,\mathrm{d}s},
\end{aligned}$$

which is the Poincaré inequality.

We now prove the equivalence of the two norms. By Hölder's inequality, we have that

$$|\mathrm{avg}(h)|^2 \leq \frac{1}{L}\int_0^L |h(s)|^2\,\mathrm{d}s$$

so that $\|h\|_{\tilde{H}^1} \leq \|h\|_{H^1}$. On the other hand, note that

$$\frac{1}{L}\int_0^L |h(s) - \mathrm{avg}(h)|^2\,\mathrm{d}s = \frac{1}{L}\int_0^L |h(s)|^2\,\mathrm{d}s - |\mathrm{avg}(h)|^2,$$

so that

$$\begin{aligned}
\|h\|_{H^1}^2 &= \int_0^L \frac{1}{L}|h(s)|^2 + \lambda L|h'(s)|^2 \,\mathrm{d}s \\
&= \frac{1}{L}\int_0^L |h(s) - \mathrm{avg}(h)|^2 \,\mathrm{d}s + \int_0^L \lambda L|h'(s)|^2 \,\mathrm{d}s + |\mathrm{avg}(h)|^2 \\
&\leq |\mathrm{avg}(h)|^2 + L(1+\lambda)\int_0^L |h'(s)|^2 \,\mathrm{d}s \leq ((1+\lambda)/\lambda)\|h\|_{\tilde{H}^1}^2
\end{aligned}$$

One can iterate Poincaré's inequality to demonstrate the equivalence of H^n and $\tilde{H}^n$.

□

The use of the Poincaré inequality can also be used to show the equivalence between the generic H^n norm (53) and the H^n norm we consider (58).

Note that we have not established any relation between the geometry of the inner products H^n and $\tilde{H}^n$; however, in the next sections, we show that the gradients with respect to H^n and $\tilde{H}^n$ have similar properties.

3.3 H^1 and $\tilde{H}^1$ Gradients

In this section, we describe how to compute first order Sobolev gradients from the H^0 gradient. Denote by $f = \nabla_{H^0} E(c)$ the gradient of E with respect to the H^0 inner product at c. We would like to compute first the H^1 gradient at c. Assuming $g = \nabla_{H^1} E(c)$ exists, we have for all $h \in T_c M$,

$$\mathrm{d}E(c)\cdot h = \langle h, g\rangle_{H^0} + \lambda L^2 \langle h', g'\rangle_{H^0} = \langle h, g - \lambda L^2 g''\rangle_{H^0}$$

where we have integrated by parts and noted that we have periodic boundary conditions. Since gradients are unique (if they exist), we must have that

$$f(s) = g(s) - \lambda L^2 g''(s) \quad \text{where } s \in [0, L]. \tag{61}$$

Note that this is an ODE defined on $[0, L]$ with periodic boundary conditions, that is, all derivatives match on the boundary of $[0, L]$.

Now we take a similar approach to compute the $\tilde{H}^1$ gradient. Assuming $g = \nabla_{\tilde{H}^1} E(c)$ exists, we have

$$\mathrm{d}E(c) \cdot h = \mathrm{avg}(h) \cdot \mathrm{avg}(g) + \lambda L^2 \langle h', g' \rangle_{H^0} = \langle h, \mathrm{avg}(g) - \lambda L^2 g'' \rangle_{H^0} .$$

Again by uniqueness, we have that $f = \mathrm{avg}(g) - \lambda L^2 g''$. Noting periodic boundary conditions, we have that $\mathrm{avg}(g) = \mathrm{avg}(f)$, and so

$$f(s) = \mathrm{avg}(f) - \lambda L^2 g''(s) \quad \text{where } s \in [0, L] \tag{62}$$

and we have periodic boundary conditions.

3.3.1 Solving the ODEs

We want to first solve the ODE (61) for g. It suffices to solve (61) with the boundary conditions $g(0) = g(L)$ and $g'(0) = g'(L)$. One can show that $g(s) = \int_0^L k_\lambda(s, \hat{s}) f(\hat{s}) \, \mathrm{d}\hat{s}$, where $k_\lambda : [0, L] \times [0, L] \to \mathbb{R}$ satisfies the following conditions for all $s, \hat{s} \in (0, L)$

$$k_\lambda(s, \hat{s}) - \lambda L^2 \frac{\partial^2 k_\lambda}{\partial s^2}(s, \hat{s}) = \delta(s - \hat{s}) \tag{63a}$$

$$k_\lambda(0, \hat{s}) = k_\lambda(L, \hat{s}); \;\; \partial_s k_\lambda(0, \hat{s}) = \partial_s k_\lambda(L, \hat{s}) \tag{63b}$$

and δ denotes the Dirac distribution. It can be shown that the solution to the previous system is $k_\lambda(s, \hat{s}) = K_\lambda(|s - \hat{s}|)$, where $K_\lambda : \mathbb{R} \to \mathbb{R}$ is given by

$$K_\lambda(s) = \frac{\cosh\left(\frac{s - \frac{L}{2}}{\sqrt{\lambda} L}\right)}{2L\sqrt{\lambda} \sinh\left(\frac{1}{2\sqrt{\lambda}}\right)}, \quad \text{for } s \in [0, L], \tag{64}$$

and K_λ is periodically extended to all of $\mathbb{R}$. We may write

$$\boxed{\nabla_{H^1} E(s) = \int_c K_\lambda(\hat{s} - s) \nabla_{H^0} E(\hat{s}) \, \mathrm{d}\hat{s} = (K_\lambda * \nabla_{H^0} E)(s)} \tag{65}$$

where the integral over c denotes any range of $\hat{s}$ that corresponds to one full period around the curve c (*e.g.* [0,L], [–L,0], [–L/2,L/2], etc.).

We now solve the second ODE (62). It suffices to solve (62) with the boundary conditions $g(0) = g(L)$, $g'(0) = g'(L)$, and the relation $\operatorname{avg}(f) = \operatorname{avg}(g)$. Integrating (62) twice yields

$$g(s) = g(0) + sg'(0) - \frac{1}{\lambda L^2} \int_0^s (s - \hat{s})(f(\hat{s}) - \operatorname{avg}(f)) \, \mathrm{d}\hat{s}. \tag{66}$$

Using (66), applying the boundary conditions, and noting that $\operatorname{avg}(g) = \operatorname{avg}(f)$, after some manipulation, yields

$$g'(0) = -\frac{1}{\lambda L^3} \int_0^L s(f(s) - \operatorname{avg}(f)) \, \mathrm{d}s \quad \text{and} \quad g(0) = \int_0^L f(s) \tilde{K}_\lambda(s) \, \mathrm{d}s \tag{67}$$

where the formula for $\tilde{K}_\lambda$ is in equation (68).

Remark 3.3.1. *We want to remark that the above yields a solution to the ODE (62)* without resorting to a convolution*; the same is true for the similar equation for the $\tilde{H}^n$ gradients. So, $\tilde{H}^n$ gradients may be computed from H^0 gradients in $O(N)$ time, where N is the number of vertices of the polygon approximation.*

The equation (62) may be nonetheless solved using a convolution, where the kernel function $\tilde{K}_\lambda$ is given by

$$\tilde{K}_\lambda(s) = \frac{1}{L} \left(1 + \frac{(s/L)^2 - (s/L) + 1/6}{2\lambda} \right), \quad s \in [0, L]. \tag{68}$$

and it is (unsurprisingly) the same kernel that is used in equation (67). Note that $\tilde{K}_\lambda(0) = \tilde{K}_\lambda(L)$ and thus we may periodically extend $\tilde{K}_\lambda$ as before. In this case, we may rewrite, $g(0) = \int_c f(\hat{s}) \tilde{K}_\lambda(\hat{s}) \, \mathrm{d}\hat{s}$, where, again, the integral over c denotes any range of $\hat{s}$ that corresponds to one full period over c. Therefore,

$$\boxed{\nabla_{\tilde{H}^1} E(s) = \int_c \tilde{K}_\lambda(\hat{s} - s) \nabla_{H^0} E(\hat{s}) \, \mathrm{d}\hat{s} = (\tilde{K}_\lambda * \nabla_{H^0} E)(s).} \tag{69}$$

3.3.2 Properties of the Kernels

Note the following formal properties of K_λ and $\tilde{K}_\lambda$:

$$K_\lambda''(s) = \frac{1}{\lambda L^2} \left(K_\lambda - \delta(s) \right) \quad \text{and} \quad \tilde{K}_\lambda''(s) = \frac{1}{\lambda L^2} \left(\frac{1}{L} - \delta(s) \right), \quad s \in [0, L). \tag{70}$$

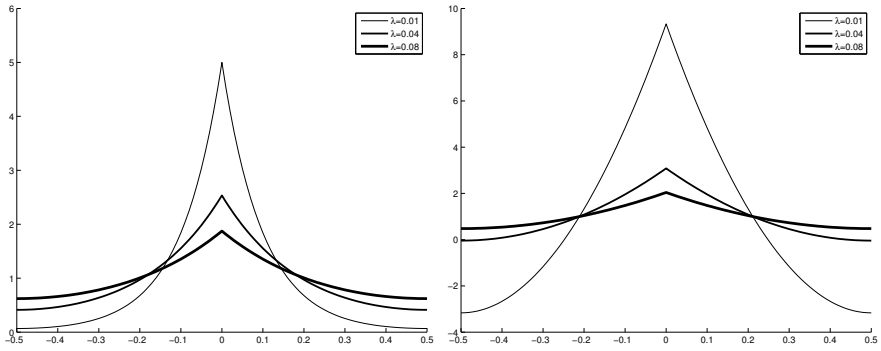


Figure 10: Plots of K_λ (left) and $\tilde{K}_\lambda$ (right) for various λ with $L = 1$. The plots show the kernels over one period.

The first property is just the relation in (63a), and the second is obtained through differentiation of $\tilde{K}_\lambda$. Using these relations, it is easy to see that $K_\lambda * f$ and $\tilde{K}_\lambda * f$ formally solve (61) and (62), respectively. Next, note that

$$\int_c K_\lambda(\hat{s})\,\mathrm{d}\hat{s} = 1 \quad \text{and} \quad \int_c \tilde{K}_\lambda(\hat{s})\,\mathrm{d}\hat{s} = 1 \tag{71}$$

for all $\lambda > 0$. Also observe that $K_\lambda \geq 0$ for all $\lambda > 0$, and that $\tilde{K}_\lambda \geq 0$ only when $\lambda \geq 1/24$. Finally, it is easy to verify that as $\lambda \to +\infty$, $K_\lambda \to 1/L$ and $\tilde{K}_\lambda \to 1/L$. See Figure 10 for plots of K_λ and $\tilde{K}_\lambda$.

3.3.3 Properties of Sobolev Gradients

First note, from formulas (65) and (69), that the H^1 and $\tilde{H}^1$ gradients are geometric, *i.e.*, they do not depend on a particular parameterization chosen for the curve. This is also evident from the definition of these inner products. The formulas (65) and (69) show that there may be a tangential component of the gradients; but these tangential components may be ignored when considering gradient flows. This is different from H^0 where if the energy is geometric, then the gradient will have only a normal component.

Because H^1 and $\tilde{H}^1$ gradients are given by integrals of the H^0 gradient, given in formulas (65) and (69), integration by parts and the relations in (70) imply that two derivatives of the curve can be moved to derivatives on the kernels. This means that H^1 and $\tilde{H}^1$ gradients involve two fewer derivatives of the curve than H^0 gradients

involve. Note that H^0 gradients have twice the number of derivatives of the curve as is defined in the energy E to be optimized. Thus, fourth order evolution equations of curves in H^0 may reduce to second order equations in H^1 and $\tilde{H}^1$. A similar remark can be made for H^n and $\tilde{H}^n$ gradients; these gradients require $2n$ less derivatives of the curve than the H^0 gradient requires.

The property that the integral of both the kernels is unity (71) implies that the H^1 gradient can be interpreted as a weighted average of the H^0 gradient; the same interpretation holds for $\tilde{H}^1$ when $\lambda > 1/24$. In light of this weighted average interpretation, we see that Sobolev gradients are less sensitive to noise and local features than H^0 gradients are. Moreover, the property that the kernels approach $1/L$ as $\lambda \to +\infty$ shows that, in this case, the H^1 and $\tilde{H}^1$ gradients approach pure translations equal to the average value of the H^0 gradient, as expected from the interpretation of gradient noted in Section 3.2.1.

While local minimizers in the H^0 metric continue to be local minimizers in the H^1 metric (but not vice-versa), the definition of locality is completely different. As a result, Sobolev active contours are much more robust to the local minima which strongly influence H^0 active contours. This property arises due to the fact that Sobolev active contours are able to exploit much more relevant information contained in the initial contour about the shape of the desired contour. This is because the Sobolev active contour is more resistant to changing the local structure of the initial curve since local perturbations induce high derivatives in the flow field compared to more global motions which are smoother. Thus, if one starts with a contour that is smooth on a local scale, the cost to perturb the initially smooth contour into a "noisy" version of the same contour is enormous. This renders a very large distance between a smooth and non-smooth contour even though the two contours may be very close in other senses, such as in the Hausdorff distance. As such, local minimizers due to noise are no longer "local" to the initial contour if the initial contour is smooth and

the local minimizer is not. Switching to the Sobolev metric has the effect of pushing many undesirable minimizers so far away from the initial contour that they are no longer able to influence the gradient flow towards a more desirable minimizer which is now much closer to the initial contour.

Remark 3.3.2. *One may wonder whether any smoothing kernel can used to smooth out the H^0-gradient to form a smoother perturbation, which is robust to noise. The answer may be yes, but it is not guaranteed to reduce the energy in question. The fact that the kernels we've derived come directly from a gradient means that the energy is guaranteed to be reduced.*

The previous comment also relates to various numerical methods for energy descent algorithms where the H^0 gradient is modified by an operator known as a pre-conditioner to yield a better descent direction for faster convergence to a solution. For example, the Newton method uses the inverse Hessian of the energy as the pre-conditioner (see for example [46, 13] for applications to active contours). The disadvantage of this approach is that one has to be careful to make sure the resulting algorithm reduces the energy; even for the Newton algorithm, the inverse Hessian pre-conditioner is not guaranteed to reduce the energy, in general.

3.3.4 Advantages of $\tilde{H}^1$ over H^1

There is a computational advantage of using the $\tilde{H}^1$ gradient as opposed to the H^1 gradient since the formulas (66), (67) give the $\tilde{H}^1$ gradient as a single integral without convolution, as opposed to the convolution for H^1. Another advantage of $\tilde{H}^1$ over H^1 is that we can eliminate the dependence on the parameter λ when implementing $\tilde{H}^1$ gradient flows. Observe from the kernel definition (68) that $\tilde{K}_\lambda$ is a sum of two terms: one that depends on λ and another that does not. Thus, the $\tilde{H}^1$ gradient is a sum of two components: one that depends on λ by a simple scale factor, and another that is independent of λ. The component that does not depend on λ is $\text{avg}(\nabla_{H^0} E)$, which is

a just a translation. The other component is a complex deformation. An algorithm to implement an approximate version of the $\tilde{H}^1$ gradient flow is to first evolve the curve by the translation component until this component becomes zero, then to evolve the curve by the deformation component, and the process is repeated until convergence. Note that λ does not need to be chosen for evolving the deformation component because λ only changes the speed of the curve, not the geometry. Therefore, this algorithm also gives a way of separating the (rigid) motion of the curve from the deformation. Separating the motion from deformation has particular importance in tracking applications [99]. Note that this algorithm is equivalent to optimizing the energy first according to the $\tilde{H}^1$ inner product as $\lambda \to +\infty$, and then according to $\tilde{H}^1$ with any λ (as there is no dependence on λ after the first step).

3.4 Some Sobolev Gradient Flows

In this section, we simplify the formulas (65) and (69) for some common geometric energies, note some interesting properties, and compare these with the usual H^0 gradients. A question that arises when considering these gradient flows is whether an initial curve $c \in M$, the manifold of curves, stays in the manifold of curves (this also relates to existence of a solution for the PDE). The manifold of curves consists of curves being immersed and regular. Since the flows we consider are geometric and we represent the curve by arclength parameterization, the curves always remain immersed by representation if their derivative is defined. The question of regularity is a difficult one in general, and we do not address it in this thesis.

In what follows, we use K to denote either the kernel (64) or (68), and ∇_1 will denote either the H^1 or $\tilde{H}^1$ gradient; when the distinction is needed, we will use the subscript λ on the kernels, and write H^1 or $\tilde{H}^1$.

3.4.1 Length and Weighted Length

We consider the geodesic active contour model [16, 53]. The energy is

$$E(c) = \int_c \phi(c(s))\,\mathrm{d}s$$

where $\phi : \mathbb{R}^2 \to \mathbb{R}^+$. Then the gradient with respect to H^0 is

$$\nabla_{H^0} E = L(\nabla\phi(c) \cdot \mathcal{N})\mathcal{N} - L\phi(c)\kappa\mathcal{N}$$

where $\mathcal{N}$ is the unit inward normal, and κ is the curvature. We omit the argument in $\phi(c)$, for simplicity, and write c' for arc parameter derivation. Let us first note that $\nabla_{H^0} E = L\nabla\phi - L(\phi c')'$. Integrating by parts we find that

$$\frac{1}{L}\nabla_1 E = \nabla\phi * K - (\phi c')' * K = \nabla\phi * K - (\phi_s c) * K' - (\phi c) * K'',$$

where $\phi_s := (\phi \circ c)'$. Using the relations in (70), we find that

$$\nabla_{\tilde{H}^1} E = \frac{\phi c - \mathrm{avg}(\phi c)}{\lambda L} - L(\phi_s c) * \tilde{K}'_\lambda + L\nabla\phi * \tilde{K}_\lambda. \tag{72}$$

The above does not require that the curve be twice differentiable, and thus we may prove that

Proposition 3.4.1. *Suppose that $\phi \in C^1$, $\phi > 0$; then the gradient flow*

$$\frac{dc}{dt} = -\nabla_{\tilde{H}^1} E(c)$$

exists for small (positive) times; if $\liminf_{x\to\infty} \phi(x)|x| > 0$, then it exists for all (positive) times.

Proof. The proof is based on the inequality

$$|\nabla_{\tilde{H}^1} E(s)| \le M_r(1+r)\frac{1}{\lambda} + LM_r r\frac{1}{2\lambda} + LM_r \quad \forall s$$

where the ball of radius r centered at 0 contains c, and

$$\frac{1}{M_r}E(c) \le L \le \frac{1}{m_r}E(c)$$

where

$$M_r := \sup_{|x|<r}(\max\{\phi(x), |\nabla\phi(x)|\}) \text{ and } m_r := \inf_{|x|<r} \phi(x) \ ;$$

the above (by using the Euler method) implies existence, in the class of curve evolutions $c(\theta, t)$ that are Lipschitz in both variables.

Details will be available in a forthcoming paper. □

Of particular interest is when $\phi = 1$, that is $E = L$, the length of the curve. We see that

$$\nabla_{\tilde{H}^1} L = \frac{c - \text{avg}(c)}{\lambda L}.$$

It is interesting to notice that the H^1 and $\tilde{H}^1$ gradient flows are stable for both ascent and descent while the H^0 gradient flow is only stable for descent. Note that the $\tilde{H}^1$ gradient flow constitutes a simple rescaling of the curve about its centroid. While the H^0 gradient descent smooths the curve, the $\tilde{H}^1$ gradient descent (or ascent) has neither a beneficial nor a detrimental effect on the regularity of the curve.

3.4.2 Area and Weighted Area

We consider region-based active contour models; for example, [114, 17]. The energy is

$$E(c) = \int_{c_{in}} \phi \, \mathrm{d}A,$$

where c_{in} denotes the region enclosed by the closed curve c, $\phi : \mathbb{R}^2 \to \mathbb{R}$ and $\mathrm{d}A$ is the area form; this energy is defined only for embedded curves. The gradient with respect to H^0 is

$$\nabla_{H^0} E = -L\phi\mathcal{N} = -L\phi Jc'$$

and J is a rotation by 90° matrix. Integrating by parts we find that

$$\frac{1}{L}\nabla_1 E = -(\phi Jc') * K = (\phi_s Jc) * K + (\phi Jc) * K'. \tag{73}$$

For the $\tilde{H}^1$ gradient, this simplifies to

$$\nabla_{\tilde{H}^1} E = \frac{J}{\lambda L^2} \int_0^L \Big(\phi c(\cdot + \hat{s}) - \text{avg}(\phi c)\Big)\hat{s} \, \mathrm{d}\hat{s} + (\phi_s Jc) * \tilde{K}_\lambda. \tag{74}$$

Of particular interest is when $\phi = 1$, that is $E = A$, the area enclosed by the curve. We see that $\nabla_1 A = (Jc) * K'$. This simplifies to the gradient ascent/descent

$$C_t(s) = \pm \frac{J}{\lambda L^2} \int_0^L \Big(C(s+\hat{s}) - \text{avg}(C) \Big) \hat{s} \, \mathrm{d}\hat{s} \tag{75}$$

in the $\tilde{H}^1$ gradient case.

3.4.3 Elastic Energy

Consider the elastic energy defined by

$$E(c) = \int_c \kappa^2(s) \, \mathrm{d}s,$$

where L is the length of c, and κ is the signed curvature. This may serve as a regularizer that does not favor smaller length curves, unlike the standard curve shortening term. It is also the term that is commonly omitted in the original snakes model of [51] since it leads to a fourth order gradient flow.

We now derive the $\tilde{H}^1$ flow by first calculating the H^0- gradient. Let us write a family of curves as $C : \mathbb{R}^+ \times S^1 \to \mathbb{R}^2$, then we may write

$$E(t) = \int_C C_{ss} \cdot C_{ss} \, \mathrm{d}s.$$

Let us first compute some intermediate formulas. Let $f : \mathbb{R}^+ \times S^1 \to \mathbb{R}$, then

$$\begin{aligned} \frac{\partial}{\partial t}\frac{\partial}{\partial s} f &= \frac{\partial}{\partial t}\frac{1}{\|C_p\|}\frac{\partial}{\partial p} f = -\frac{C_{tp} \cdot C_p}{\|C_p\|^3}\frac{\partial}{\partial p} f + \frac{1}{\|C_p\|}\frac{\partial}{\partial p}\frac{\partial}{\partial t} f \\ &= -C_{ts} \cdot C_s \frac{\partial}{\partial s} f + \frac{\partial}{\partial s}\frac{\partial}{\partial t} f, \end{aligned}$$

or more conveniently,

$$f_{st} = f_{ts} - (C_{ts} \cdot C_s) f_s. \tag{76}$$

We now compute

$$\frac{\partial}{\partial t}(C_{ss} \cdot C_{ss}) = 2 C_{sst} \cdot C_{ss},$$

but using (76) twice, we see

$$\begin{aligned} C_{sst} &= C_{sts} - (C_{ts} \cdot C_s)C_{ss} \\ &= \frac{\partial}{\partial s}(C_{ts} - (C_{ts} \cdot C_s)C_s) - (C_{ts} \cdot C_s)C_{ss} \\ &= C_{tss} - (C_{tss} \cdot C_s + C_{ts} \cdot C_{ss})C_s - (C_{ts} \cdot C_s)C_{ss} - (C_{ts} \cdot C_s)C_{ss} \\ &= C_{tss} - (C_{tss} \cdot C_s + C_{ts} \cdot C_{ss})C_s - 2(C_{ts} \cdot C_s)C_{ss}, \end{aligned}$$

and since $C_s \cdot C_{ss} = 0$, we have

$$\frac{\partial}{\partial t}(C_{ss} \cdot C_{ss}) = 2(C_{tss} \cdot C_{ss}) - 4(C_{ts} \cdot C_s)(C_{ss} \cdot C_{ss}). \tag{77}$$

Now,

$$\begin{aligned} E'(t) &= \frac{\mathrm{d}}{\mathrm{d}t}\int_0^1 C_{ss} \cdot C_{ss}\|C_p\| \,\mathrm{d}p \\ &= \int_0^1 \frac{\partial}{\partial t}(\|C_p\|)C_{ss} \cdot C_{ss} \,\mathrm{d}p + \int_C \frac{\partial}{\partial t}(C_{ss} \cdot C_{ss}) \,\mathrm{d}s. \end{aligned}$$

By substituting (77), we have

$$\begin{aligned} E'(t) &= \int_C (C_{ts} \cdot C_s)(C_{ss} \cdot C_{ss}) \,\mathrm{d}s \\ &+ 2\int_C (C_{tss} \cdot C_{ss}) - 2(C_{ts} \cdot C_s)(C_{ss} \cdot C_{ss}) \,\mathrm{d}s \\ &= -\int_C 3(C_{ts} \cdot C_s)(C_{ss} \cdot C_{ss}) \,\mathrm{d}s + 2\int_C (C_{tss} \cdot C_{ss}) \,\mathrm{d}s \\ &= \int_C (2(C_{tss} \cdot C_{ss}) - 3(C_{ts} \cdot C_s)(C_{ss} \cdot C_{ss})) \,\mathrm{d}s. \end{aligned}$$

Integrating by parts, we find

$$E'(t) = \frac{1}{L}\int_C C_t \cdot (2L\partial_{ss}(c_{ss}) + 3L\partial_s((c_{ss} \cdot c_{ss})c_s)) \,\mathrm{d}s.$$

Hence,

$$\nabla_{H^0} E(c) = 2Lc_{ssss} + 3L\partial_s((c_{ss} \cdot c_{ss})c_s). \tag{78}$$

Computing the Sobolev gradient from (78), we have

$$\begin{aligned} \nabla_{\tilde{H}^1} E &= K * \nabla_{H^0} E \\ &= 2LK'' * (c_{ss}) - 3LK' * ((c_{ss} \cdot c_{ss})c_s) \end{aligned}$$

Hence, for the $\tilde{K}_\lambda$ kernel we have

$$\nabla_{\tilde{H}^1} E = -\frac{2}{\lambda L}\kappa\mathcal{N} - 3LK' * (\kappa^2\mathcal{T}) \tag{79}$$

since $\text{avg}(\kappa\mathcal{N}) = 0$. Notice that the corresponding $\tilde{H}^1$ gradient flow is second order, although it is an integral-PDE.

3.4.4 Comparison of H^0 and H^1, $\tilde{H}^1$

We notice several advantages of the gradients flows for H^1 and $\tilde{H}^1$ gradients as compared with H^0 gradients. First note that both the expressions for edge-based and region-based active contour gradients with respect to H^1 and $\tilde{H}^1$ (72), (73) do not involve any derivatives of the curve. This is in contrast to H^0, which requires two derivatives for geodesic active contours and one derivative for region-based active contours. Hence, these Sobolev flows are defined for non-smooth curves, *e.g.*, polygons, without the need to resort to numerical techniques based on viscosity solutions of the corresponding level set equations. Moreover, Proposition 3.4.1 shows that we do not need viscosity theory to define the $\tilde{H}^1$ flow at least for the weighted length energy. Note that the expression in (72) does not require any more derivatives of ϕ than the expression for H^0 does. This is not the case for (73), which requires a derivative of ϕ. However, since ϕ_s is contained within a convolution, the possible noise generated by ϕ_s is mitigated. Alternatively, the original expressions (65) and (69) may be used if a derivative of ϕ is not desired to be computed.

Notice the expressions of Sobolev gradients for the elastic energy (79) only require two derivatives of the curve; this is in contrast to the H^0 gradient, which requires four derivatives of the curve. Since there is no maximum principle for fourth order equations, the H^0 gradient descent of the elastic energy is difficult to implement using level set methods [23, 31]. A particle method can be used; however, this is prone to numerical problems. Note that the integral-PDE (79) may not have a maximum principle, but in our numerical implementation with level set methods (see Section 3.5.2),

we bypass this issue for the following reason. The local term certainly has a maximum principle, and for the global term, we perform *extensions* of this quantity from the zero level set, which is done, for example, in image-based evolutions. Therefore, we expect that level sets do not collide, that the level set function gradient doesn't become ill-defined, and that the level set evolution is modeling the curve evolution. These are issues that need to be dealt with for the H^0 gradient flow using a level set method.

3.5 Numerical Implementation

In this section, we describe the numerical scheme used to simulate Sobolev active contours. Sobolev active contours are naturally suited for a parametric or a marker particle based implementation. This is because computing Sobolev gradients requires computing an integral around the curve, which is straightforward to compute if one has an ordered set of sample points of the curve.

3.5.1 Multiple Curves Evolution

In this section, we consider evolving multiple curves according to gradient flows for the metrics defined in Section 3.2.2. We consider defining the flows for curves that undergo topological changes. Consider a family of embedded curves $c = (c_i)$ in the plane and of displacements $h = (h_i)$, where $c_i, h_i : S^1 \to \mathbb{R}^d$ and $i = 1 \dots N = N(c)$. We may define a metric on multiple curves $\langle , \rangle_c$ using a choice of some metric $\langle , \rangle_{c_i}$ that was defined for a single curve in Section 3.2.2. We define the inner product on multiple curves to be

$$\langle h, k \rangle_c := \sum_i \langle h, k \rangle_{c_i}.$$

Consider an energy, E, on multiple curves defined by

$$E(c) := \sum_{i=0}^{N(c)} E(c_i),$$

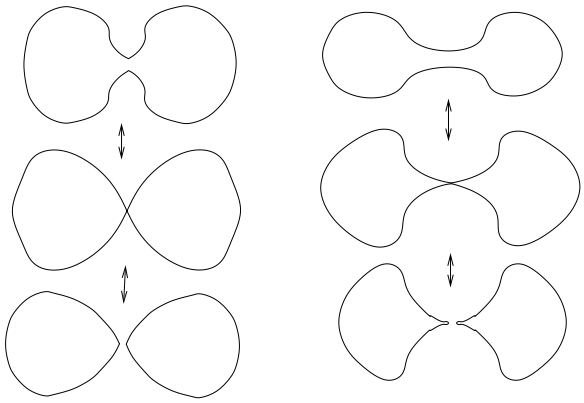


Figure 11: Some illustrations of topological changes.

then, the variation of E for several embedded curves is

$$\mathrm{d}E(c) \cdot h = \sum_{i=0}^{N(c)} \langle \nabla E(c_i), h_i \rangle_{c_i} ,$$

for some given inner product. Thus, we consider evolving the curves according to the curve evolution

$$\partial_t c_i = -\nabla E(c_i) \quad i = 1, \ldots, N(c). \tag{80}$$

Clearly, if the curves are moved according to the above equation, and the curves do not undergo topological changes then the evolution reduces the energy, E.

We now consider the possibility of topological changes of c, that is merging and splitting of curves in c. Two ways of topological changes are pictorially shown in Figure 11. On the left, the curves develop kinks; on the right, they develop cusps. Note that other topological changes may be possible. We consider whether the evolution of (80) reduces the energy, E of the curve c, even when topology changes occur. First, we note a problem for actually defining the evolution in (80) for Sobolev metrics. Note that during the instant in which the curve undergoes a topology change an ambiguity arises in how to parameterize the curve at the point of self-intersection. Since the Sobolev inner product depends upon the arclength parameterization of the curve this leads to an ambiguity in defining the gradient and the resulting curve evolution. For example, on the right side of Figure 11, at the instant of topology change, there

is an ambiguity over whether to treat c as a single curve or as two separate curves. In the case of the H^0 gradient, this ambiguity is unimportant since in both cases the H^0 gradient is the same at all points of the curve except the point of self-intersection as the H^0 does not depend upon the curve globally as does the Sobolev gradient. Since gradient descent flows are designed to minimize their corresponding energy functional as quickly as possible, there is a very natural solution to the ambiguity that arises in the case of the Sobolev gradient at the moment a topological change occurs. In theory, one may simply compute the derivative of the energy corresponding to either interpretation of the curve (each of which results in a different gradient under the Sobolev metric) and choose the one which the larger derivative as it will lead to a faster instantaneous descent of the energy. In practice, though, an even simpler numerical procedure may be employed based on the same essential idea. We will outline the procedure below for the case of level set methods.

When employing level set methods for Sobolev active contours (as described in Section 3.5.2), the contour extracted at any discrete moment in time from the discrete spatial grid always exhibits the simple topology of an embedded curve(s), which bypasses the ambiguity discussed above. Thus a topological change occurs more abruptly after one discrete evolution step in which the curve passes from one topology to another, never passing through the ambiguous intermediate configuration in which the curve no longer remains embedded. Thus, there is never a need to resolve the Sobolev gradient of the curve in such configurations. Instead, noting that there is typically a discontinuity in the Sobolev gradient flow before and after topology change, we may simply verify that the energy has decreased after the topological change. If E is continuous, there is no reason to expect repeated oscillations in the topology of the curve even though the Sobolev gradient flow may exhibit a discontinuity across the topology change. The weighted length and area energies defined in Section 3.4 satisfy this continuity condition (since these functionals remain unambiguous through

a topology change even though their Sobolev gradients do not). Indeed on the right-hand side of Figure 11, these energies remain constant for the curve just before and just after the topology change. On the other hand, energies that depend on local derivatives of the curve such as the elastic energy defined in Section 3.4 may not be continuous through the topology change. As such there is no guarantee that the energy E decreases after an evolution step that creates a topological change. If the energy decreases, it makes sense to continue evolving. If instead an increase is detected, then one should stop evolving the curve and restore its prior configuration. Otherwise, oscillations may occur.

While the elastic energy pose a potential problem for Sobolev metrics (and indeed even the H^0 metric) in the event of certain types of topological changes (those for which the elastic energy increases), they pose an even greater problem for the standard H^0 metric which yields gradient flows of fourth order that are much more difficult to implement numerically even in the absence of topological changes.

3.5.2 Level Set Method

We first describe a straightforward numerical method to extend curve evolutions that depend on integrals around the curve to level set evolutions, introduced in [78]. The algorithm for updating the level set function, $\Psi : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}$, for most Sobolev active contours is

1. Compute polygon(s) estimate of zero level set of Ψ from the narrowband, U_n.
2. Calculate and interpolate H^0-gradient to polygon(s) estimate
3. Compute Sobolev gradient on polygon(s) estimate by using one of the formulas for the H^1 or $\tilde{H}^1$-gradients in terms of the H^0-gradient shown in Section 3.3.1.
4. Extend polygon(s) forces to narrowband band of level set function domain.

5. Evolve Ψ by the transport equation $\Psi_t = -\nabla\Psi \cdot \vec{F}$, where $\vec{F} : U_n \subset \mathbb{R}^2 \to \mathbb{R}$ is the extended Sobolev gradient to the narrowband region, U_n.

The computational complexity of the entire algorithm for extracting the polygon(s), and computing the extensions is linear in the size of the narrowband, U_n.

In most of the cases of various energy functionals, the H^0 gradient of the energy must be computed on the narrowband U_n in the usual fashion; or in our case, we directly compute the H^0-gradient to the polygon estimate of $\{\Psi = 0\}$. In other cases (such as for the elastic energy), the H^0 gradient does not need to be computed, but only certain expressions that do not involve the convolutions. For example, in the elastic energy, the term $\kappa^2\mathcal{T}$ should be computed at every point of the polygon estimate. Note that in computing quantities such as the normal vector, and curvature of $\{\Psi = 0\}$, we compute them directly from Ψ using standard formulas and interpolate them to the polygon estimate. The Sobolev gradient is then computed using a convolution or simple integral with the formulas in Section 3.3.1. Note that we must compute separate convolutions or integrals on each polygon extracted from Ψ (or connected component of Ψ) in the fashion described in Section 3.5.1. We use a standard Riemann sum to compute the curve integrals, and the polygon estimate is used to determine the arclength measure, $\mathrm{d}s$. After the Sobolev gradient is computed on the polygon estimate, it can be extended to U_n. We call the extended Sobolev gradient $\vec{F}$. Note that we are extending the gradient along the curve so that $\nabla\Psi \cdot \nabla F_i = 0$ where $i = 1, 2$ and $\vec{F} = (F_1, F_2)$. That is, we assume that all the level sets of Ψ are moving by the same speed function, which means that the level sets of Ψ do not collide. Now the level set evolution can be computed by

$$\Psi_t(x) = -\nabla\Psi(x) \cdot \vec{F}(x).$$

3.5.3 Level Set Method Without Polygon Extraction

In this section, we describe an approach where we can directly compute (without polygon extraction) an approximation to the Sobolev gradient from the H^0 gradient that is defined on the narrowband, U_n, of the level set function. This may be of use since the polygon extraction step is cumbersome.

Our approach is to approximate the convolution integrals (65) and (69) with region integrals evaluated within the narrowband U_n. The arc-distance between points on the curve (*i.e.*, $s - \hat{s}$) that is required by the formulas (65) and (69) can be approximated by the use of the eikonal equation [90]. Consider an embedded curve, $c \in C^2(S^1, \mathbb{R}^2)$, and the function $F : \mathbb{R}^2 \to \mathbb{R}^+$ defined as

$$F(x) = \begin{cases} +\infty & \text{for } x \notin \{c\} \\ 1 & \text{for } x \in \{c\} \end{cases}$$

where $\{c\} = \{c(\theta) : \theta \in S^1\}$. Define $u : \mathbb{R}^2 \to \mathbb{R}^+$ by

$$u(c(s)) := \inf_{\gamma \in \Gamma} \int_0^1 F(\gamma(\theta)) \|\gamma'(\theta)\| \, \mathrm{d}\theta,$$

where $\Gamma := \{\gamma : [0,1] \to \mathbb{R}^2, \gamma(0) = c(\hat{s}), \gamma(1) = c(s)\}$; then clearly, we see that $u(c(s))$ equals the minimum distance between the points $c(s)$ and $c(\hat{s})$ along the curve c. Because of the symmetry of the kernels in (64) and (68), it suffices to use the quantity $u(c(s))$ as a substitute for $s - \hat{s}$ in the convolutions formulas (65) and (69). As in [90], we may solve for the viscosity solution of the eikonal equation

$$\|\nabla u(x)\| = F(x), \quad u(c(\hat{s})) = 0 \tag{81}$$

to obtain the desired solution of u at all points along the curve, c. As we will solve this equation numerically on a grid, we consider the following approximation to (81),

$$\|\nabla u(x)\| = 1 + |\Psi(x)|/\epsilon, \quad u(c(\hat{s})) = 0 \tag{82}$$

where $\epsilon > 0$ is chosen small enough, and Ψ is the level set function with $\Psi(c(s)) = 0$ for all s.

Next, note the co-area formula for a Borel measurable function $f : \mathbb{R}^2 \to \mathbb{R}$, a Borel set $A \subset \mathbb{R}^2$ such that $\nabla\Psi(x) \neq 0$ for $x \in A$,

$$\int_A f(x)\|\nabla\Psi(x)\| \, \mathrm{d}x = \int_{\Psi(A)} \int_{\Psi^{-1}(t)} f(x) \, \mathrm{d}\mathscr{H}^1(x) \, \mathrm{d}t$$

where $\mathrm{d}\mathscr{H}^1$ is the one-dimensional Hausdorff measure, that is, arc-length measure. We use this formula to convert the contours integrals of interest to integrals over the domain of the level set (or the narrowband region). Thus, we find by the co-area formula and the Lebesgue differentiation theorem,

$$\int_c H(c(s))K(\hat{s} - s) \, \mathrm{d}s \approx \int_A H(x)K(u(x))\delta_a(\Psi(x))\|\nabla\Psi(x)\| \, \mathrm{d}x$$

where $\delta_a : \mathbb{R} \to \mathbb{R}$ is a smooth approximation to the Dirac distribution, $H : \mathbb{R}^2 \to \mathbb{R}$ is some function, u is the solution to (82), and $A \subset \mathbb{R}^2$, such that $\{c\} \subset A^o$, the interior of A. According to the previous comments, we propose the following theorem:

Theorem 3.5.1. *Suppose that $c : S^1 \to \mathbb{R}^2$ is a C^2 curve embedded in the plane, Ψ is zero only on the image of c, and $\nabla\Psi(c(\theta)) \neq 0$ for $\theta \in S^1$; then,*

$$\lim_{\epsilon \to 0} \frac{1}{2\epsilon} \int_A |\nabla\Psi(x)| Z(|\Psi(x)|/\epsilon) \, \mathrm{d}x = \textit{length of the curve} \tag{83}$$

where $Z : \mathbb{R}^+ \to \mathbb{R}^+$ is a positive continuous decreasing functions s.t. $Z(0) = \int_0^\infty Z = 1$ and $\lim_{x\to\infty} xZ(x) = 0$ (for example $Z(x) = \exp(-x)$); the integral on the left hand side is on a compact set A such that the curve is contained in A^o, the interior of A.

Fix a continuous $H : \mathbb{R}^2 \to \mathbb{R}$, a kernel $K : \mathbb{R}^+ \to \mathbb{R}$, and suppose that $\lim_{x\to\infty} K(x)/x = 1$. Let $\epsilon > 0$, and $c(\bar{s})$ be a point in the curve and let $u : \mathbb{R}^2 \to \mathbb{R}$ be the viscosity solution of

$$|\nabla u(x)| - (1 + |\Psi(x)|/\epsilon) = 0 \quad , \quad u(c(\hat{s})) = 0 \tag{84}$$

then

$$\lim_{\epsilon \to 0} \frac{1}{2\epsilon^2} \int_A |\nabla\Psi(x)| \, H(x) \, K(u(x)) \, Z(|\Psi(x)|/\epsilon^2) \, \mathrm{d}x = \int_c H(c(s)) \, K(|s - \hat{s}|) \, \mathrm{d}s \tag{85}$$

where $|s - \hat{s}|$ is the shortest arclength distance between $c(s)$ and $c(\hat{s})$ along c.

To solve the eikonal equation (84) numerically, we use the fast marching method [2]. We discretize on a $N \times N$ grid, then the integrals above are numerically substituted by a sum; if $\epsilon \to 0$ while N is kept constant, then clearly the numerical versions of (83) and (85) would converge to zero rather than to the desired result; so we propose the following numerical schemes for the approximation of (83) and (83):

$$\frac{1}{2\epsilon N^2} \sum_{i,j}^{N} |\nabla \Psi(x_{i,j})| Z(|\Psi(x_{i,j})|/\epsilon) \approx \text{length of the curve} \tag{86}$$

where $x_{i,j}$ are points of the discretization of the domain of Ψ to an $N \times N$ grid, and $\epsilon \to 0$, $N \to +\infty$ and $N\epsilon \to +\infty$. Similarly, we use the method by [2] to solve for a numerical approximation $u_{i,j}^N$ to (84); then (85) becomes

$$\begin{aligned} \frac{1}{2\epsilon^2 N^2} \sum_{i,j}^{N} |\nabla \Psi(x_{i,j})|\, H(x_{i,j})\, K(u_{i,j}^N)\, Z(|\Psi(x_{i,j})|/\epsilon^2)\, \mathrm{d}x \approx \\ \int_c H(c(s))\, K(|s - \bar{s}|)\, \mathrm{d}s. \end{aligned} \tag{87}$$

3.6 Experiments

In this section, we show some simulations of Sobolev active contours used for segmentation of static images and some simple image sequences.

We tried both the method of Section 3.5.2 and 3.5.3; the latter is simpler to implement, but is difficult to tune w.r.t. the choice of ϵ and N. The following numerical experiments use the polygon(s) estimate of the zero level set, as in Section 3.5.2.

In all the simulations done below, the results for the Sobolev active contours are done with the $\tilde{H}^1$ inner product ($\lambda = 10$ unless stated otherwise, although a wide range of λ give similar results). Using the H^1 inner product gives visually similar results as to what are shown. We consider two energies to illustrate the advantages of Sobolev active contours over H^0 active contours. The edge-based energy we consider

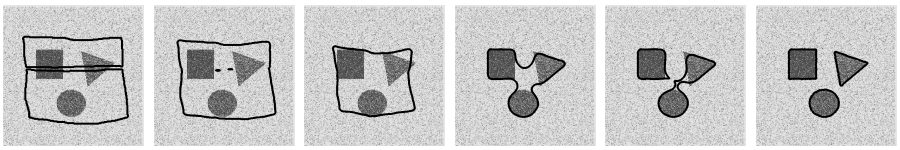


Figure 12: Segmentation of various shapes using a Sobolev active contour for a region-based energy (89). This illustrates the ability of Sobolev active contours to cope with topological changes: merging and splitting.

is

$$E_e(c) = \int_c \phi(c(s))\,\mathrm{d}s, \text{ where } \phi = \frac{1}{1+\|\nabla I\|^2}, \tag{88}$$

which was proposed by [16, 53]. The region-based energy we consider is

$$E_r(c) = \int_{c_{in}} (I-u)^2\,\mathrm{d}A + \int_{c_{out}} (I-v)^2\,\mathrm{d}A + \alpha L(c), \tag{89}$$

where

$$u = \frac{\int_{c_{in}} I\,\mathrm{d}A}{\int_{c_{in}} \mathrm{d}A}, \text{ and } v = \frac{\int_{c_{out}} I\,\mathrm{d}A}{\int_{c_{out}} \mathrm{d}A},$$

and $\alpha \geq 0$ specifies a penalty on the length, $L(\cdot)$, of the curve. This is the piecewise-constant model of [72] (see also [17]). The use of the length penalty is partly to keep the evolving H^0 contour smooth, and avoid undesirable local minimum of the first terms caused by noise. This term has a smoothing effect since the H^0 gradient flow of L is curvature flow, which has smoothing properties; as we saw in Section 3.4.1 this length penalty for the Sobolev flow is futile in terms of giving smoothness to the contour.

3.6.1 Merging and Splitting

In Figure 12, we demonstrate the experimental evidence for the ideas presented in Section 3.5.1. In this experiment, we segment an image using the region-based energy (89), and a Sobolev active contour. We see that the active contour can change topology to achieve the global minimum.

3.6.2 Noisy Square Segmentation

In Figure 13, we show the results of an experiment in which we segment a noisy square image with the region-based energy (89). We added salt and pepper noise with densities of $\rho = 0.5, 0.6, 0.7$ to the binary image. In this experiment, we compare the results obtained from using the usual H^0 active contour with the result obtained from using a Sobolev active contour.

First, we explore the effects of various weights, α, on the length penalty in (89) for the H^0 active contour. In Figure 13, we see that with using a small α, the contour becomes stuck in the noise, at an intermediate local minimum of E_r (89). The image second from the right in each row is the result of the H^0-active contour with a minimum α just high enough so that the contour is not stuck in noise at a local minimum. In Figure 13(a), we see that with relatively low noise, the result of the H^0-active contour with just high enough α to overcome the local minimum captures the desired square accurately. As the noise increases, a higher α is needed to overcome local minima; however, because of such a high length penalty and therefore higher smoothness, the active contour is unable to capture the fine-scale structure of the desired object. In Figure 13(c), we see that using the minimal α to overcome the noise results in the contour failing to capture the corners of the desired square object.

Notice that for each of the noise levels, the Sobolev active contour without any length penalty ($\alpha = 0$) or additional regularization terms in the energy moves in such a way as to avoid any local minimum of the energy, E_r, as it moves in a global fashion first before resorting to finer scale deformations (the final segmentation of the square images are the last image in each row of Figure 13). The end result captures the square accurately and looks to be the global minimum of the region-based energy. The result is independent of λ, the weighting on the derivative component of the $\tilde{H}^1$ inner product since translations do not optimize the energy. Notice that the converged Sobolev contour becomes more rugged as the noise increases. This is

because the image is corrupted by the noise, and the original square is no longer the global minimum of energy. We should point out that the Sobolev active contour gives smoothness in the contour *flow*; it does not guarantee smoothness in the final contour, which is determined by the energy that is being minimized.

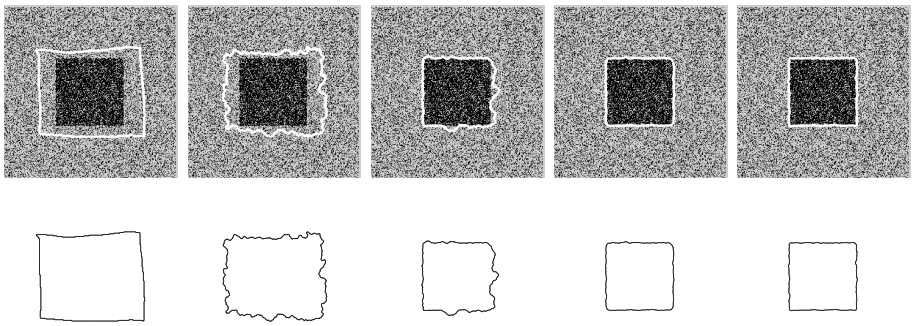
Besides obtaining a more accurate segmentation, the Sobolev active contour converges much faster than the H^0 active contour with just high enough length penalty, α, to get passed local minima. The reason for this is because of the small step size needed for stability of the curvature term arising from the length penalty in the level set implementation. A crude estimate for the Sobolev active contour shows that the step size is less than $0.5/255^2 = 130\,050^{-1}$, whereas for the H^0-active contour with $\alpha = 300\,000$, as in Figure 13(c), a step size of less than $\min\{0.5/300\,000 = 600\,000^{-1}, 0.5/255^2 = 130\,050^{-1}\} = 600\,000^{-1}$ is needed. This is a factor of about 5 for this crude estimate.

3.6.3 Segmentation of Real Images

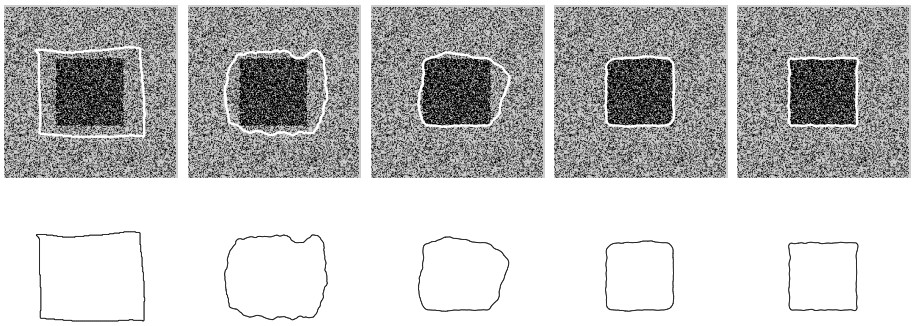
We illustrate the advantages of using Sobolev active contours over H^0 with the same energy on real medical images in Figures 14, 15, and 16.

In Figure 14, we use the edge-based energy, E_e (88) to segment a cardiac image. We show the results with two different initializations using both H^0 and Sobolev active contours. Because of the complicated texture of the image, the edge-based energy is riddled with local minima. Because the Sobolev active contour moves more globally, it is less likely than the H^0 active contour to become stuck in local minima, as shown in Figure 14.

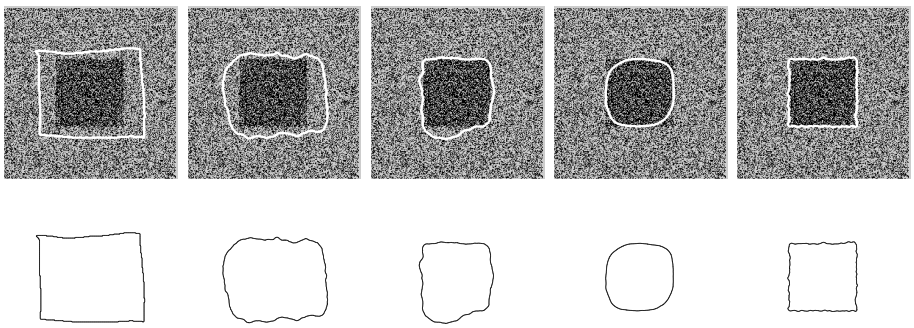
In Figure 15, we segment an ultrasound image using the region-based energy (89). In Figure 15(a), the results are shown using an H^0 active contour with two different weightings, α, on the length penalty. In both cases, extraneous features of the image are detected. In Figure 15(b), we see that the Sobolev contour moves



(a) $\rho = 0.5$. Left to right: Initialization, $H^0 : \alpha = 20\,000, 50\,000, 90\,000$, Sobolev.

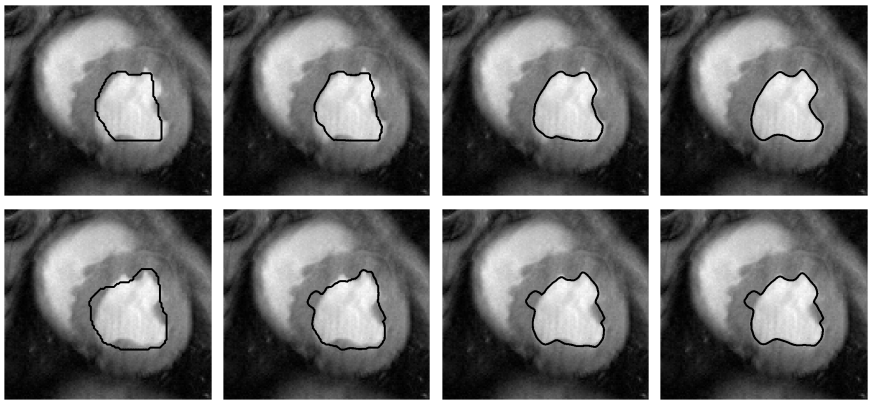


(b) $\rho = 0.6$. Left to right: Initialization, $H^0 : \alpha = 50\,000, 120\,000, 160\,000$, Sobolev.

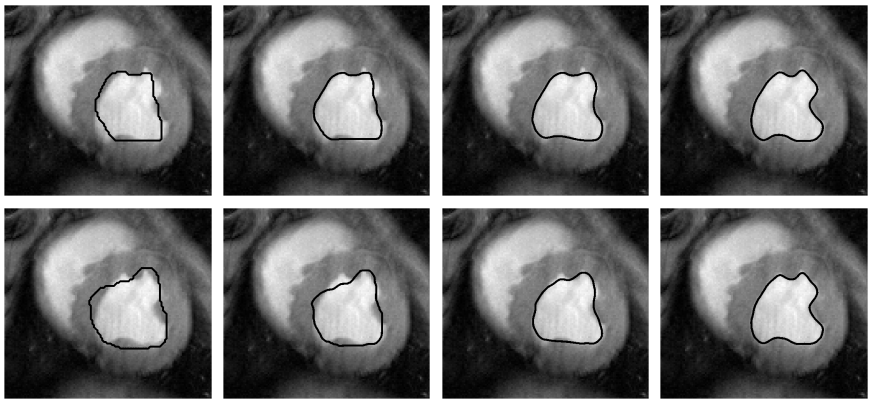


(c) $\rho = 0.7$. Left to right: Initialization, $H^0 : \alpha = 50\,000, 100\,000, 300\,000$, Sobolev.

Figure 13: Segmentation of square binary image with salt and pepper noise of various densities. The experiment shows the results with H^0 (of various degrees of regularization, α) and the Sobolev active contour.

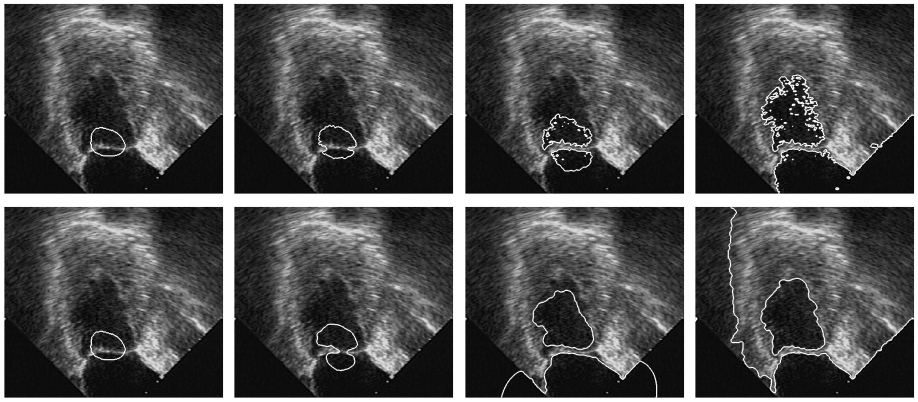


(a) H^0 active contour.

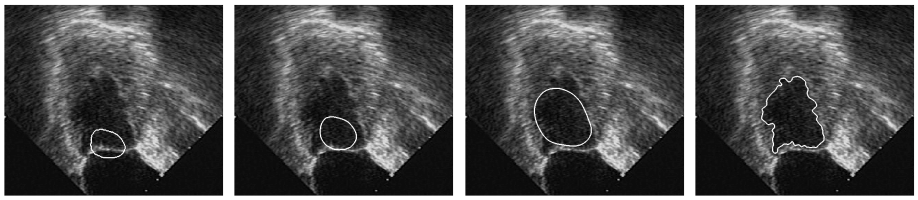


(b) Sobolev active contour.

Figure 14: Segmentation of a cardiac image using the edge-based energy in (88) (two different initializations).



(a) H^0 active contour with low ($\alpha = 10$, top) and with high ($\alpha = 5000$, bottom) regularization.



(b) Sobolev active contour (no regularization).

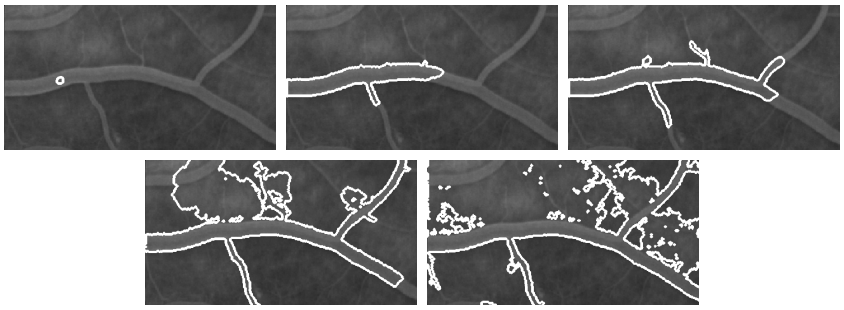
Figure 15: Segmentation of an ultrasound image using the region-based energy (89).

initially according to global motions (translation and dilation), and finally detects the more fine features of the image when more global motions cannot optimize the energy. Thus, the contour is able to avoid irrelevant local features that disturbs the H^0 active contour.

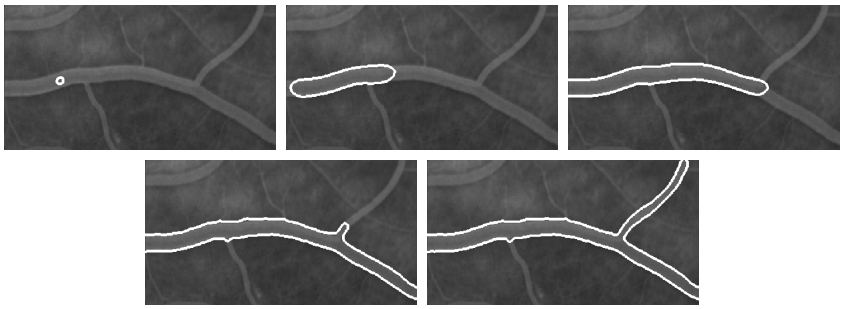
In Figure 16, we segment a vessel image using a the region-based energy (89). As we see, the Sobolev active contour is less affected by local features, which causes the H^0 active contour to leak into an irrelevant region of the image.

3.7 Conclusion

In summary, we have observed that much of the literature on active contours uses the concept of *gradient flow* to minimize energies, but it has always been assumed (knowingly or unknowingly) that the inner product on curve perturbations, on which the gradient depends, is the H^0 inner product. We have introduced using Sobolev



(a) H^0 active contour with length penalty.



(b) Sobolev active contour (no length penalty).

Figure 16: Segmentation of a vessel image using the region-based energy (89) with both the H^0 active contour and the Sobolev active contour. The Sobolev active contour is able to avoid distracting fine features of the image and therefore does not leak into an irrelevant region of the image.

inner products on the set of perturbations of a curve. We have demonstrated the general methodology for computing Sobolev gradients, which requires integrating the H^0 gradient. The procedure requires very little extra computational time and little change of existing computer code when compared with the H^0 gradient flow, in particular for the $\tilde{H}^n$ family of norms, as remarked in 3.3.1. It was demonstrated by theory and experiments that Sobolev gradient flows are global flows, in which a single point on the curve depends on all other points of the curve, and the flows deform locally after global motions (*e.g.*, translations) can no longer optimize the energy. This particular property shows that the Sobolev method gives a smooth *flow* (not necessarily a smooth contour), which in many cases helps avoid certain undesirable local minimum of active contour energies that disturb the (local) H^0 flows. Explicit formulas for Sobolev gradient flows of typical energies found in the active contour literature, which were derived, showed many interesting properties of Sobolev active contours. One of the notable properties that these explicit formulas shows is that the Sobolev method regularizes the corresponding H^0 gradient flow by reducing the order of the PDE. In particular, derivatives of the curve need not be defined for region-based and edge-based energies, and the elastic energy, which is results in a fourth order H^0 flow is reduced to a second order flow using a first order Sobolev flow.

CHAPTER IV

COARSE-TO-FINE SEGMENTATION AND TRACKING USING SOBOLEV ACTIVE CONTOURS

Sobolev active contours introduced in Chapter 3 introduced a new paradigm for minimizing energies defined on curves by changing the traditional cost of perturbing a curve and thereby redefining their gradients. Sobolev active contours evolve more globally and are less attracted to certain intermediate local minima than traditional active contours, and it is based on a well-structured Riemannian metric, which is important for shape analysis and shape priors. In this chapter, we analyze Sobolev active contours using scale-space analysis in order to understand their evolution across different scales. This analysis shows an extremely important and useful behavior of Sobolev contours, namely, that they move successively from coarse to increasingly finer scale motions in a continuous manner. This property illustrates that one justification for using the Sobolev technique is for applications where coarse-scale deformations are preferred over fine scale deformations. Along with other properties to be discussed, the coarse-to-fine observation reveals that Sobolev active contours are, in particular, ideally suited for tracking algorithms that use active contours. We will also justify our assertion that the Sobolev metric should be used over the traditional metric for active contours in tracking problems by experimentally showing how a variety of active contour based tracking methods can be significantly improved merely by evolving the active contour according to the Sobolev method.

4.1 Introduction

Tracking objects in video sequences with active contours has been an active research area ever since the introduction of *snakes* in [51] (see [6] for a survey). This is often a two step procedure. The first step is *detection*. Here an initial estimate of the object boundary being tracked in a particular image (video frame) is given, and the goal is to evolve this initial contour towards the object of interest in that particular frame. A wide variety of different energy-based schemes have been proposed to do this, including both edge-based [51, 16, 53, 39] and region-based [72, 118, 114, 79, 17, 80] active contours. The second step is *prediction*, where the objective is to predict the object's boundary in the upcoming image based on the presently detected contour as well as contours detected in previous images. Measured (or assumed) dynamics are then extrapolated forward to predict the upcoming contour. Many times, the result from prediction is then averaged in an appropriate manner with the result of detection on the prediction to form an *estimate* of the contour (for example see [104, 49]). A trivial approach, which we call the *naive tracker*, assumes no change in dynamics and therefore uses the contour detected in the current frame as the prediction (initial contour) for the next frame. More sophisticated prediction steps may be found in [5, 104, 48, 82] for parametric snakes and more recently [75, 49, 84] for geometric active contours. In [49], a dynamical model of the object being tracked is used, and [84] uses particle filtering for geometric active contours building on the approach in [48].

The prediction step in many contour tracking algorithms is needed because the detection step is too sensitive to initial contour placement, thereby rendering the *naive tracker* inadequate. Indeed, if we had a robust detection scheme that could operate in real-time, then the prediction step could be eliminated and the *naive tracker* would suffice. This sensitivity of active contour models comes in part due to a lack of inherent smoothness in the way the active contours evolve or deform.

Typically an object being tracked deforms rather smoothly from frame to frame, otherwise a prediction would make no sense. Note that we are referring to smoothness of the contour *deformation* (along the contour), not the contour itself. Active contour energies, through the use of regularizers, may easily be adapted to favor smoothness in the final detected contour. However, in tracking it makes sense to ensure smoothness of the deformation of the contour from one frame to the next, regardless of how smooth we want the contour to be. Most current and previous active contour algorithms allow an initial contour to deform in very complex ways, as it flows toward an energy minimum. Even if the final contour has the exact same shape as the initial contour up to translation, the intermediate contours attained during the evolution may vary immensely from the initial and final shapes. This non-preferential *freedom* of the contour to undergo arbitrarily complicated deformations as it flows can attract the contour to undesirable, intermediate local minima before it reaches the desired object boundary.

It would thus be beneficial, when tracking with active contours, to evolve the initial contour, whether or not it was obtained by the naive tracker or by a prediction step, toward its final configuration in a manner that mimics the evolution behavior of objects we wish to track. In particular, it would be ideal if the evolution first favored rigid motions that did not change the actual shape of the evolving contour and then gave preferential treatment to coarser or more global deformations, resorting only at the end to finer or more local deformations when necessary.

In Chapter 3, *Sobolev Active Contours* introduced a new paradigm for minimizing energies defined on curves (see also [19, 21]). This yields a completely new way to evolve active contours by exploiting the fact that the gradient flow used to evolve a contour is not only influenced by the energy it minimizes but also by how we measure the cost of perturbing the curve. The works [68, 113] revealed many undesirable properties associated with the usual cost (H^0) inherent in all previous geometric active

contour models. Accordingly, in Chapter 3, we considered using other norms for perturbing active contours based on Sobolev spaces. Sobolev active contours evolve more globally and are less attracted to certain intermediate local minima than traditional active contours. In contrast to the usual strategy of substituting simple energies with more complex (and costly) energies exhibiting fewer local minima, Sobolev active contours minimize the same energy, but follow an entirely different deformation to reach their steady state configuration, thereby avoiding many local minima that would otherwise have been encountered along the way.

Applying Sobolev norms to variational problems has been done in areas other than active contours to gain many of the same advantages that are gained in active contour problems. For example, the book [74] (see also references within) presents the theory of Sobolev gradients and applies it to various physical problems. The key difference between those metrics and the ones we are considering is that we are defining *geometrized* Sobolev norms; these are to be thought of as the metrics for a (yet to be completely studied) Riemannian manifold of curves, equivalent up to reparametrization – whereas the usual Sobolev norms are associated to Hilbert (or Banach) vector spaces of *parametrized* curves. Recent work that uses the Sobolev gradient for boundary reconstruction is [12]. Geometrized Sobolev metrics have been used for defining shape spaces for *shape analysis*, where the objective is to be able to perform statistical operations on *shapes*, which may be contours. For example, [117, 67, 69] consider Sobolev metrics on the space of plane contours for shape analysis, and other variants are considered by [70].

The contribution of this chapter is to expand Sobolev active contours to tracking applications and give a detailed new analysis, which among other things explains the compelling reasons for using the Sobolev metric in tracking situations. Indeed, we examine Sobolev active contours using a scale-space type analysis which shows, along with other properties to be discussed, that these active contours are quite naturally

suited for tracking problems, performing (given the exact same energy functional) significantly better than the corresponding traditional active contour. This makes the generic tracking algorithm less dependent on its prediction step as the initial contour does not need to be placed within as narrow an attraction basin in order to reach the desired minimum. In fact, we will see Sobolev active contours often allows even the *naive tracker* to perform well with simple energies that are otherwise plagued by undesirable local minima problems. For a more detailed discussion and analysis of the benefits of Sobolev active contours for tracking, see Section 4.4.

4.1.1 Related Works

We now discuss previous techniques that have been explicitly designed to obtain multiscale and global motion properties that are naturally inherent in Sobolev active contours, discuss the advantages of Sobolev active contours in relation to these techniques.

Many active contour works have explicitly incorporated information from successive *scales* of an image to perform a systematic segmentation that matches image data at both coarse and fine scales. For example, in [59] the image is down-sampled to a coarse scale, and an active contour [51] is evolved until convergence. The resulting active contour is up-sampled to a finer scale of the image, and the process is continued on successively finer scale representations of the image until the active contour is evolved in the image itself. The method of [59] makes it less likely that the active contour becomes stuck in irrelevant local minimum of the underlying energy caused by fine scale features of the image. There is also a computational advantage of these methods since the algorithm works with down-sampled images.

- One problem with this method is knowing which and how many scales of the image to use. Ideally, one would like to use a continuum of scales that have gradually more information added at each successive scale. However, this is

practically not possible in the framework of [59].

- When an arbitrary discrete sampling of scales is chosen, there is a greater possibility of the evolving contour becoming trapped in artificial local minima.

- Choosing a large number of scales reduces the chance of being trapped in a local minimum; however, the computational cost increases.

- One limitation of multiscale techniques like [59] is that there is a limit to how much the domain of the image may be coarsened whenever complicated geometrical objects are present in the image, and it is not trivial to obtain the limit. To illustrate this point, consider a simple example of a binary image consisting of two large circles connected by a long thin strip (an elongated version of the object in Fig. 22). If coarsened or down-sampled enough, the object will become two circles separated by a space, a completely different topology than the actual object. Imagine that an active contour is initialized to enclose the two circles, and the geodesic active contour energy [16, 53] (see also (88)) is used. The contour will split and become two circles. When the contour is evolved in a finer scale of the image, which shows the strip, the contour will not capture the strip because there is a more favorable local minimum of the energy, which is the curve consisting of two circles that slightly protrude into the strip, but do not capture it. In general, this problem is exhibited whenever thin structures are present in the image. This example serves to illustrate that pathologies arising from topological inconsistency of the object to be segmented *across scale* can result from multiscale methods like [59] if careful attention is not paid to the amount of coarsification performed. We note that Sobolev active contours has no such limitation in that *the object to be segmented* stays a consistent topology *across each instant of time* during the contour evolution (even when complicated geometrical objects and even thin structures are present).

A solution to the problem of selecting scales is proposed in [8], which considers segmenting all scales of a scale space of an image simultaneously. Although this method has some advantages, the main drawback is a significant increase in computational time because a surface rather than a curve evolution is needed.

As shown in the Chapter 3, Sobolev active contours are *global* flows in that they incorporate image and curve information from the entire curve in order to evolve a single point on the curve. Typical multiscale methods that are used for curve evolutions also use global information in order to deform the contour, but the notion of "global" is different than for Sobolev active contours. Indeed in the coarse scale evolution of a multiscale algorithm, a point $c(s)$ of the curve uses image information from the neighborhood $\{x \in \mathbb{R}^2 : \|x - c(s)\| < R\}$ where $R > 0$ depends on the amount of smoothing/down-sampling performed. While this gives some advantages over traditional active contours including that the evolution incorporates coarse scale information before finer details, the evolution of the curve is not coarse-to-fine and there is no preference to coarse scale deformations before finer deformations because the traditional metric is used. Our scale-space type analysis shows that Sobolev active contours has inherent multiscale behavior. Indeed, the Sobolev technique incorporates information from *all* scales of the image in a systematic fashion first incorporating coarse scale information and favoring coarse deformations of the curve, before gradually moving to finer deformations when coarse deformations can no longer minimize the energy of interest.

Numerous approaches have been applied to tracking that deal with the problem of traditional active contours being attracted to irrelevant local features of an image resulting from the flexibility of the contour to undergo arbitrary deformations. The approach taken by some to avoid arbitrary (and unlikely) deformations in tracking applications has been to restrict the degrees of freedom of the active contour so that more global deformations of the contour are only possible. For example, this approach

has been taken by [28, 108, 109]. In [28], a deforming contour is represented by b-splines, and [108, 109] uses polygons. Since there are fewer control points for a b-spline and polygon than a typical parametric snake, this results in more global deformations of the contour. An advantage of [108, 109] over traditional active contours is that information is *integrated* over adjacent edges of the polygon in order to move the corresponding vertex, and hence this adds robustness of the polygon to noise and other local features. Other related methods use of a *finite* number of Fourier or wavelet coefficients to represent the evolving contour (for example, see [59]).

- An advantage of these methods that restrict the contour to more global deformations, besides being less attracted to local features, is that they are computationally fast compared to both parametric and level set implementations [78]. This is because they have inherent regularization and typically do not require an additional regularization term (such as a curvature term), which implies that these methods may take much larger time steps than traditional methods.

- One disadvantage is that topological changes become hard to handle (see [66]).

- Another disadvantage of these approaches is that the motion of the contour is restricted, and therefore, it becomes impossible to detect fine features of the image when they are needed.

- Moreover, these methods do not generally evolve in a coarse-to-fine fashion, which is an advantage of the Sobolev active contours.

Techniques have been designed to force a coarse-to-fine evolution of active contours in order to avoid undesirable local minima of energies in tracking applications. For example, in [115, 99], the authors propose to optimize energies that are defined on both the set of curves and on a set of *global* group motions. In the simplest case, these energies are defined as $E_{\text{new}}(c, g) := E(g \circ c)$ where E is an active contour

energy, c is a curve and g is a global group action (e.g. affine motion). For tracking, it is beneficial to optimize with respect to g first while keeping c fixed since the global motion is the most important, and then to optimize (using the traditional metric) with respect to c to obtain fine scale changes of the object being tracked. More recently, [19, 21] (see a related idea in [64]) have proposed to optimize energies using "spatially coherent" flows to achieve the effect of [115, 99] by constructing inner products on the space of perturbations of a curve that favor various group motions such as affine motions. The corresponding norms are equivalent (unlike Sobolev-type norms) to the traditional H^0 norm, and thus the notion of "locality" is the same for both H^0 and "spatially coherent" norms (see Section 4.2.2 for a consequence of this fact). This approach is useful for tracking if one has a prior assumption that the object of interest is moving according to an affine motion. The advantage of using Sobolev active contours for tracking is the inherent coarse-to-fine behavior, which the inner products based on group motions do not have, and the fact that explicit groups do not need to be chosen. In many tracking situations (for example see the real sequences in Section 4.5), it is not necessarily the case that the object of interest is moving according to an affine motion nor close to it (w.r.t. boundary movements). However, one can generally say that the object is moving according to "coarse" deformations w.r.t. its boundary that cannot be captured with simple groups motions (such as affine). The Sobolev active contour is ideal for such a situation where one does not have an explicit representation of the motion because it favors general coarse-scale deformations before finer deformations.

4.2 Some Motivation for Sobolev Metrics

4.2.1 Consistent Theory of Shape

One of the main motivations of our work on Sobolev metrics in the space of curves is to obtain a *consistent* way of doing both shape analysis (*e.g.*, computing statistics

of shapes) and shape optimization (*e.g.*, optimizing active contour energies) in a *consistent* manner, and to date this is the only work that offers such a consistent theory. Indeed, it has recently been proven that the Riemannian metric arising from the Sobolev metrics we consider in the space of curves yield well-defined, that is non-zero distances, and that the metrics are complete with respect to Lipschitz curves [67]. Also recently, [69] computed the geodesic equation for the H^n metric, and proved the existence of geodesics for small times and smooth initial data. In the past, there have been many metrics proposed for doing shape analysis (e.g. [107, 20, 99, 70]), but the optimization procedure for energies defined on the space of shapes is completely inconsistent with this metric and the geometry of the space, and therefore completely artificial.

There are many advantages of the Riemannian metric approach for shape analysis, for example being able to define a principal component analysis of a set of shapes: we may perform a PCA in the tangent space to the mean shape on the vectors that point in the geodesic direction of each individual shape. A consistent approach for optimization on the space of curves is desired for many reasons. For example, consider the simple application of incorporating prior information into segmentation using active contours (e.g. see [60, 107, 89, 22, 26]). In the simplest case, we may have prior information that the object to be detected from an image is close in the sense of our metric to the shape c_0. Our active contour energy may be the following:

$$E(c) = E_{\text{image}}(c) + d(c, c_0)$$

where c is a shape (e.g. curve), d is the Riemannian metric (shape metric) on the space of curves, and E_{image} is an image-based term. To minimize the energy, one can consider calculating the gradient descent flow, which depends on the metric used to define gradient. If we choose consistent Riemannian metrics for d and the gradient, then the gradient of $d(\cdot, c_0)$ becomes simply the vector pointing in the direction of the geodesic, which is quite natural. On the other hand, if we make an inconsistent

choice, then the gradient of $d(\cdot, c_0)$ is the direction which maximizes the change in $d(c, c_0)$ while also minimizing the (inconsistent) cost of perturbing c, which is quite artificial. Moreover, when minimizing E with respect to a different metric than d, we end up far away (in the sense of d) from the initial curve as we step in the gradient direction, which may have detrimental effects for tracking applications where the initial curve is usually quite close to the target curve. Even when computing the average of two shapes, which is a typical computation for shape analysis, a typical procedure is to use a gradient descent to find the average. While using the same metric to define the gradient leads to a gradient descent that is intimately tied to geodesics and therefore to the geometry of the shape space, an inconsistent metric has no such natural connection to the geometry of the space.

4.2.2 Fewer Numerical Local Minimizers

A very important motivation for using Sobolev metrics to define gradient flows is the fact that, as mentioned eariler, the notion of "local" in the space of curves changes. Indeed since the Sobolev norms are not equivalent to the H^0 norm, the notion of "local" completely changes. Local minimizers for the same energy in H^0 may not be local minimizers when considering Sobolev norms (they may change to critical points that are not local minimizers). However, a local minimizer wrt Sobolev norms is a local minimizer wrt H^0 since Sobolev norms dominate the H^0 norm (i.e., $\|h\|_{H^0} \leq \|h\|_{\text{Sobolev}}$). Another consequence of the change of locality in Sobolev norms is that numerically many local minimizers of an energy due to noise vanish. Indeed curves that are local to a local minimizer due to noise in H^0 are "pushed" so far away from the "local" minimizer when considering Sobolev metrics that numerically the local minimum nolonger exists. These previous points are not true for the "spatially coherent" norms where rigid motions are favored (considered in [21]) since these norms are equivalent to the H^0 norm.

To demonstrate the numerical vanishing of local minima due to noise, we conduct the following experiment:

1. Initialize contour in a noisy image.

2. Run H^0 gradient flow on the energy

$$E(c) = E_{cv}(c) + L(c), \tag{90}$$

   where E_{cv} is the Chan-Vese energy, and L is the length functional. Call the converged contour c_0.

3. Adjust c_0 at **one sample point** by **one pixel**; call modification $\hat{c}_0$ ($\hat{c}_0$ is an H^0 local perturbation of c_0).

4. Run and compare H^0, translation favored H^0 and Sobolev gradient flows initialized with $\hat{c}_0$.

The results are given in Fig. 17. The H^0 active contour flows back to the local minimizer, c_0, since that is local to $\hat{c}_0$ wrt to H^0. The same is true for the translation favored H^0 flow since the corresponding norm is equivalent to the H^0 norm. On the otherhand, the Sobolev active contour is not attacted to the "local" minimizer c_0 since $\hat{c}_0$ is no longer local to c_0. *Note that identical results are achieved for this experiment even if the local minimum c_0 is not perturbed!!* This is because numerically a local minimum cannot be represented exactly due to finite precision issues, and these finite precision perturbations are local wrt the H^0 norm but not wrt Sobolev norms. Therefore **numerically, these local minimum due to noise no longer exist**.

4.3 Scale-Space Type Analysis of Sobolev Active Contours

In this section, through Fourier analysis, we show that Sobolev active contours favor more coarse-scale motions than regular active contours, and that they generally first undergo coarse scale motions before resorting to fine scale deformations in optimizing the chosen energy.

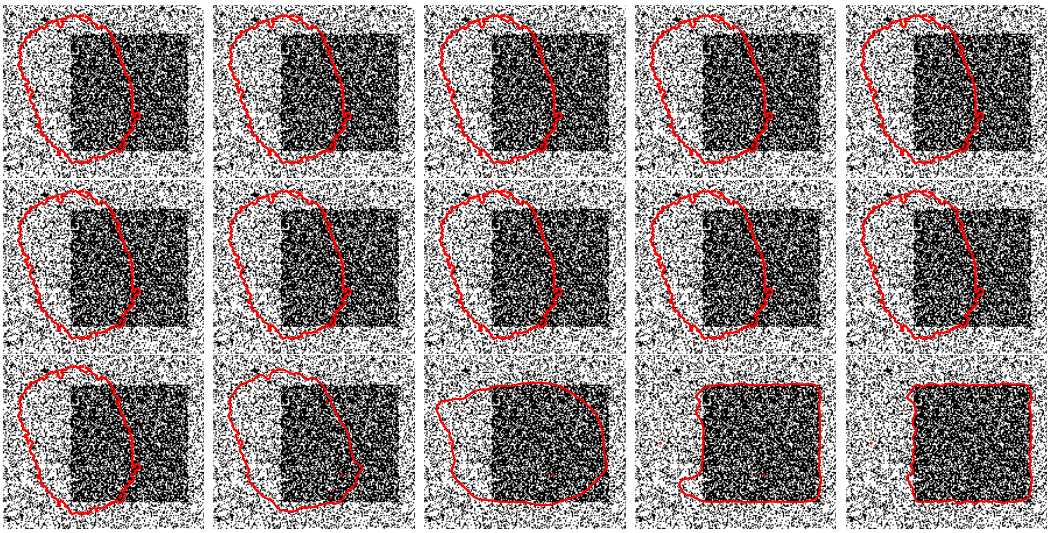


Figure 17: Top row: H^0, middle: translation favored H^0, and bottom: Sobolev flows for (90). The initialization is a "local" minimum wrt to H^0 of the energy (90) (this is $\hat{c}_0$ as noted in the text). Noise: salt and pepper (density = 0.6).

4.3.1 Sobolev Norms in Frequency Domain

Notice that since any $h \in T_c M$ is smooth on S^1, it follows that $h \in L^2(S^1)$. Thus, we may write h as a Fourier series, i.e.,

$$h(s) = \sum_{l \in \mathbb{Z}} \widehat{h}(l) \exp\left(\frac{2\pi i}{L} l s\right) \tag{91}$$

with convergence in $L^2(S^1)$ (and in fact point wise since h is smooth) where $\widehat{h} \in \ell^2(\mathbb{Z})$ is defined by

$$\widehat{h}(l) = \frac{1}{L} \int_c h(s) \exp\left(-\frac{2\pi i}{L} l s\right) \mathrm{d}s. \tag{92}$$

It should be noted that (91) decomposes the perturbation into the orthonormal basis of exponentials. This allows us to write Definition 3.2.1 in the frequency domain. By Parseval's theorem,

$$\int_0^L h(s) \cdot k(s) \, \mathrm{d}s = L \sum_{l \in \mathbb{Z}} \widehat{h}(l) \cdot \overline{\widehat{k}(l)}.$$

where $\bar{\cdot}$ denotes complex conjugation. We also have that

$$\int_0^L h^{(n)}(s) \cdot k^{(n)}(s) \, \mathrm{d}s = L \sum_{l \in \mathbb{Z}} \left(\frac{2\pi}{L} l\right)^{2n} \widehat{h}(l) \cdot \overline{\widehat{k}(l)};$$

therefore,

Proposition 4.3.1. *If $h, k \in T_cM$, L is the length of c, and $\widehat{h}, \widehat{k} : \mathbb{Z} \to \mathbb{C}$ are defined by* (92)*. Then,*

$$\langle h, k \rangle_{H^n} = \sum_{l \in \mathbb{Z}} (1 + \lambda(2\pi l)^{2n}) \widehat{h}(l) \cdot \overline{\widehat{k}(l)} \tag{93}$$

$$\langle h, k \rangle_{\tilde{H}^n} = \widehat{h}(0) \cdot \overline{\widehat{k}(0)} + \sum_{l \in \mathbb{Z}} \lambda(2\pi l)^{2n} \widehat{h}(l) \cdot \overline{\widehat{k}(l)}. \tag{94}$$

and the corresponding norms are

$$\|h\|^2_{H^n} = \sum_{l \in \mathbb{Z}} (1 + \lambda(2\pi l)^{2n}) |\widehat{h}(l)|^2 \tag{95}$$

$$\|h\|^2_{\tilde{H}^n} = |\widehat{h}(0)|^2 + \sum_{l \in \mathbb{Z}} \lambda(2\pi l)^{2n} |\widehat{h}(l)|^2. \tag{96}$$

Notice that Proposition 4.3.1 allows us to define the H^n and $\tilde{H}^n$ inner products for n that is any real number greater than 0. These inner products are defined the same way as in (93) and (94). It is easy to verify in this case too that the definitions are indeed inner products. Unfortunately for n that is not an integer, the inner products (therefore, norms) are not local, that is, they cannot be written as integrals of derivatives of the curves; but, given $r \in \mathbb{R}^+$ we can represent them, for n integer $n > r + 1/4$ as $\langle h, k \rangle_{\tilde{H}^r} = \text{avg}(h) \cdot \text{avg}(k) + L^{2n} \left\langle h^{(n)}, K * k^{(n)} \right\rangle_{H^0}$, for a kernel K with Fourier coefficients

$$\widehat{K}(l) = \begin{cases} 1 & \text{if } l = 0 \\ \lambda(2\pi l)^{2r-2n} & \text{if } l \neq 0. \end{cases}$$

The norms shown in (95) and (96) measure the perturbation magnitude in terms of its Fourier coefficients, which are the weights of its corresponding frequency components. We see that for both H^n and $\tilde{H}^n$ norms, high frequency components of the perturbation contribute increasingly to the norm of the perturbation.

4.3.2 Sobolev Gradients in Frequency Domain

We now calculate Sobolev gradients of an arbitrary energy in the frequency domain. By Definition 3.2.2, if the H^0 and H^n gradients of an energy $E : M \to \mathbb{R}$ exist, then

it follows that

$$\mathrm{d}E(c)\cdot h = \left\langle \nabla_{H^0} E(c), h\right\rangle_{H^0} = \left\langle \nabla_{H^n} E(c), h\right\rangle_{H^n}$$

for all $h \in T_cM$. Using Parseval's Theorem, the last expression becomes

$$\sum_{l\in\mathbb{Z}} (1+\lambda(2\pi l)^{2n})\widehat{\nabla_{H^n} E}(l)\cdot \widehat{h}(l) = \sum_{l\in\mathbb{Z}} \widehat{\nabla_{H^0} E}(l)\cdot \widehat{h}(l).$$

Since the last expression holds for all $h \in T_cM$, we have

$$\widehat{\nabla_{H^n} E}(l) = (1+\lambda(2\pi l)^{2n})^{-1}\widehat{\nabla_{H^0} E}(l) \quad \text{for } l \in \mathbb{Z} \tag{97}$$

Using a similar argument, we see that

$$\widehat{\nabla_{\tilde{H}^n} E}(l) = \begin{cases} \widehat{\nabla_{H^0} E}(0) & l = 0 \\ (\lambda(2\pi l)^{2n})^{-1}\widehat{\nabla_{H^0} E}(l) & l \in \mathbb{Z}\backslash\{0\}. \end{cases}$$

It is clear from the previous expressions that high frequency components of $\nabla_{H^0} E$ are less pronounced in the various forms of Sobolev gradients when compared with the H^0 gradient, with higher order Sobolev gradients damping high frequency components with faster decay rates.

4.3.3 Coarse-To-Fine Motion of Sobolev Contours

We now discuss the implications of the results in the previous sections. Note that the Fourier basis of the perturbations of a curve decomposes T_cM from global perturbations (low frequency perturbations) to increasingly more finer perturbations (high frequency perturbations). Indeed the zero frequency perturbation is a simple translation of the curve, which is completely global. See Fig. 18. Therefore, by (97), and comments in the previous section, it is apparent that Sobolev gradients yield perturbations with more pronounced global components than the standard H^0 gradient. While H^0 gradients give equal weighting across all scales, Sobolev gradients give less weight to finer scales. However, this does not mean that very fine scale deformations of the curve are restricted in Sobolev gradient flows. It just means that if there exists

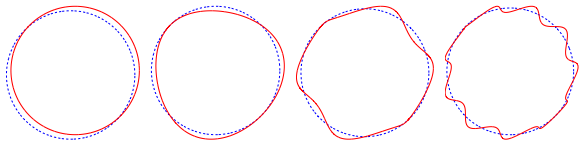


Figure 18: Increasingly higher frequency perturbations applied to a circle (left to right, $l = 0, 2, 5, 10$).

a low order perturbation (a more global motion) that increases the given energy just as would a higher order perturbation (a more finer deformation), then the low order perturbation will be preferred in the Sobolev gradient, as shown by Proposition 3.2.2. Also, if no perturbations in G_m, given by

$$G_m = \left\{ \sum_{|l| \leq m} a_l \exp\left(\frac{2\pi i}{L} l \cdot\right) : a_l \in \mathbb{C}, a_{-l} = \overline{a_l} \right\},$$

can increase the energy, E; that is $\mathrm{d}E(c) \cdot h \leq 0$ for all $h \in G_m$, then by Definition 3.2.2, we must have that $\widehat{\nabla_{H^0} E}(l) = 0$ for $l \leq m$, and therefore, we can write

$$\widehat{\nabla_{\tilde{H}^n} E}(l) = \frac{1}{\lambda (m+1)^{2n}} \begin{cases} 0, & |l| \leq m \\ \frac{1}{(2\pi l/(m+1))^{2n}} \widehat{\nabla_{H^0} E(l)}, & |l| > m \end{cases}.$$

We see that since the gradient flow does not geometrically depend on a scale factor, the Sobolev gradient automatically has the weights on high order perturbations of the gradient readjusted (so that perturbations near $|l| = m + 1$ become more pronounced). This means the Sobolev gradient flow at this particular instant of the evolution changes the finer scale structure of the curve. Thus, with Sobolev active contours, this implies at least locally, a progression from coarse scale motion to finer scale motion, unlike the standard H^0 active contour. These comments are illustrated in Fig. 19, which shows the tracking of a noisy square image using both H^0 and H^1 active contours. Notice that with the H^0 active contour, the fine structure of the curve is changed immediately, while the H^1 active contour gradually changes finer scale features of the curve after changing coarse-scale features.

The effect of using higher order (n large) Sobolev gradients is higher favorability

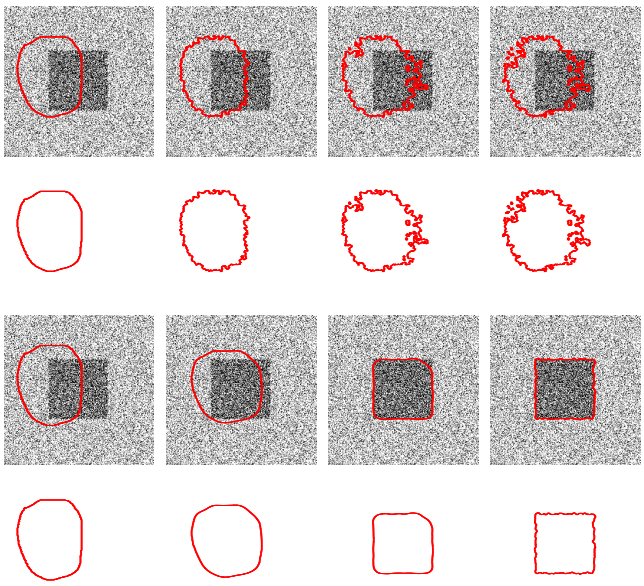


Figure 19: Standard (H^0) active contour (top) alters fine structure of the curve immediately; Sobolev (H^1) active contour (bottom) moves from coarse to finer scale motions. Both use same energy.

to lower order perturbations in the flow.

4.3.4 Analytic Examples of Coarse-To-Fine Motion of Sobolev Flows

In this section, we give two analytic examples of Sobolev gradient flows and show explicitly that as the artificial time parameter of the evolution increases, the gradient generally moves from coarse scale to finer scale perturbations. This will verify that the curve initially deforms in a coarse manner before resorting to finer deformations to decrease the energy of interest. Since geometric energies make use of the arc-length parameterization, which lead to non-linear equations that are difficult to analyze analytically, we perform an analysis on related parametric energies. Indeed, it took decades to analyze the H^0 gradient flow of a simple geometric energy, length, analytically in the mathematical community [37, 40]. However, the parametric analysis we perform is relevant in inferring the qualitative behaviors, particularly the coarse-to-fine properties, of the geometric Sobolev gradient flows as we shall justify throughout

this section by illustrating the similarities between the parametric and geometric evolutions in the cases where the geometric behavior is known through analysis from other methods.

The inner products that we use are the parametric equivalent of those given in Definition 3.2.1; that is, if $h, k : S^1 \to \mathbb{R}^2$ are perturbations of a curve $c : S^1 \to \mathbb{R}^2$ then we define

$$\langle h, k \rangle_{H^0} := \int_{S^1} h(u) \cdot k(u) \, \mathrm{d}u$$
$$\langle h, k \rangle_{\tilde{H}^n} := \mathrm{avg}(h) \cdot \mathrm{avg}(k) + \lambda \left\langle h^{(n)}, k^{(n)} \right\rangle_{H^0}$$

where $\mathrm{avg}(h) := \int_{S^1} h(u) \, \mathrm{d}u$, and $h^{(n)}$ is the n^{th} derivative with respect to the parameter u. Note that we use $\tilde{H}^n$ instead of H^n for simplicity, but similar conclusions hold for the H^n inner product

The first energy that we consider is the following curve matching energy, $E_m : C^\infty(S^1, \mathbb{R}^2) \to \mathbb{R}^+$:

$$E_m(c) = \frac{1}{2} \int_{S^1} |c(u) - c_0(u)|^2 \, \mathrm{d}u \tag{98}$$

where $c_0 : S^1 \to \mathbb{R}^2$ is a pre-specified target curve, which is the known *data*. This energy is a representative of data-based energies, e.g., possibly energies that depend on image data such as the image-based term of the Chan and Vese energy [17].

Remark 4.3.1. *A geometric version of the energy* (98) *might be the following:*

$$E(c) = \inf_{b \in [0, L_0]} \frac{1}{2} \int_c |c(s) - c_0(L_0/L_c \cdot s + b)|^2 \, \mathrm{d}s$$

where s is the arc-length parameter of c, L_0 is the length of c_0, and L_c is the length of c. An optimization of this energy would require a joint evolution of the parameter b and the curve c. The H^0 gradient of this energy w.r.t c is

$$\nabla_{H^0} E(c)(s) = (c(s) - c_0(L_0/L_c \cdot s + b)) \cdot \mathcal{N}(s) - |c(s) - c_0(L_0/L_c \cdot s + b)|^2 c_{ss}(s).$$

As we will see, the first term is the normal component of the evolution of E_m, and we will show later when we consider the energy E_s below that the second term in the

Sobolev domain has the coarse-to-fine property we seek to show. Thus, our parametric analysis gives us a good idea of the behavior of the geometric energy also.

Obviously the optimum curve for (98) is $c = c_0$, which is the global minimizer. However, we are interested in *how* the curve evolves in the frequency domain to attain the minimum. To see this, let us note our definition of Fourier transform:

$$\widehat{f}(k) = \int_{S^1} f(u) \exp(-2\pi i k u)\, \mathrm{d}u$$

where $f : S^1 \to \mathbb{R}$. Note that

$$\mathrm{d}E_m(c) \cdot h = \int_{S^1} (c(u) - c_0(u)) \cdot h(u)\, \mathrm{d}u,$$

and so, $\nabla_{H^0} E(c) = c - c_0$. Note that

$$\widehat{\nabla_{\tilde{H}^n} E_m}(l) = \begin{cases} \widehat{\nabla_{H^0} E_m}(0), & l = 0 \\ \frac{\widehat{\nabla_{H^0} E_m}(l)}{\lambda(2\pi l)^{2n}}, & l \in \mathbb{Z}\backslash\{0\}. \end{cases}$$

Let's now consider the $\tilde{H}^n$ gradient flow

$$\partial_t c(u,t) = -\nabla_{\tilde{H}^n} E_m(c);$$

we may write c as a Fourier series:

$$\begin{aligned} c(u,t) &= \sum_{l\in\mathbb{Z}} \widehat{c}(l,t) \exp(2\pi i l u) \\ \nabla_{\tilde{H}^n} E_m(u,t) &= \widehat{\nabla_{H^0} E_m}(0,t) + \sum_{l\in\mathbb{Z}\backslash\{0\}} \frac{\widehat{\nabla_{H^0} E_m}(l,t)}{\lambda(2\pi l)^{2n}} \exp(2\pi i l u) \\ &= \widehat{c}(0,t) - \widehat{c_0}(0) + \sum_{l\in\mathbb{Z}\backslash\{0\}} \frac{\widehat{c}(l,t) - \widehat{c_0}(l)}{\lambda(2\pi l)^{2n}} \exp(2\pi i l u). \end{aligned}$$

Assuming that $\widehat{c}(\cdot,t)$ is uniformly bounded by an $\ell^1(\mathbb{Z})$ function, the flow becomes

$$\sum_{l\in\mathbb{Z}} \partial_t \widehat{c}(l,t) \exp(2\pi i l u) = -(\widehat{c}(0,t) - \widehat{c_0}(0)) - \sum_{l\in\mathbb{Z}\backslash\{0\}} \frac{\widehat{c}(l,t) - \widehat{c_0}(l)}{\lambda(2\pi l)^{2n}} \exp(2\pi i l u).$$

Since $\{\exp(2\pi il\cdot)\}_{l\in\mathbb{Z}}$ is an orthogonal basis; we have that

$$\begin{aligned}\partial_t\widehat{c}(0,t) &= -(\widehat{c}(0,t)-\widehat{c_0}(0)) \quad \text{and}\\ \partial_t\widehat{c}(l,t) &= -\frac{\widehat{c}(l,t)-\widehat{c_0}(l)}{\lambda(2\pi l)^{2n}} \quad \text{for } l\in\mathbb{Z}\backslash\{0\}.\end{aligned}$$

or

$$\begin{aligned}\partial_t\widehat{c}(0,t)+\widehat{c}(0,t) &= \widehat{c_0}(0) \quad \text{and}\\ \partial_t\widehat{c}(l,t)+\frac{1}{\lambda(2\pi l)^{2n}}\widehat{c}(l,t) &= \frac{\widehat{c_0}(l)}{\lambda(2\pi l)^{2n}} \quad \text{for } l\in\mathbb{Z}\backslash\{0\}\end{aligned}$$

solving the previous equations yield

$$\begin{aligned}\widehat{c}(0,t) &= \exp(-t)\widehat{c}(0,0)+\widehat{c_0}(0)\left[1-\exp(-t)\right]\\ \widehat{c}(l,t) &= \exp\left(-\frac{t}{\lambda(2\pi l)^{2n}}\right)\widehat{c}(l,0)+\widehat{c_0}(l)\left[1-\exp\left(-\frac{t}{\lambda(2\pi l)^{2n}}\right)\right].\end{aligned}$$

Therefore,

$$\begin{aligned}\widehat{\nabla_{\tilde{H}^n}E_m}(0,t) &= \widehat{c}(0,t)-\widehat{c_0}(0) = \exp(-t)(\widehat{c}(0,0)-\widehat{c_0}(0))\\ \widehat{\nabla_{\tilde{H}^n}E_m}(l,t) &= \frac{\widehat{c}(l,t)-\widehat{c_0}(l)}{\lambda(2\pi l)^{2n}} = \frac{1}{\lambda(2\pi l)^{2n}}\exp\left(-\frac{t}{\lambda(2\pi l)^{2n}}\right)(\widehat{c}(l,0)-\widehat{c_0}(l)).\end{aligned}$$

For convenience define,

$$g_n(l,t) := \frac{1}{\lambda(2\pi l)^{2n}}\exp\left(-\frac{t}{\lambda(2\pi l)^{2n}}\right).$$

Note that

- $\lim_{t\to+\infty} g_n(l,t) = 0$ for all $l\in\mathbb{Z}$, and also the rate of convergence to zero as $|l|$ becomes larger is slower (when $n>0$). Thus, we find that the coarse scale frequency components ($|l|$ small) of c converge to the global minimum much faster than the fine scale frequencies.

- $\lim_{|l|\to+\infty} g_n(l,t) = 0$.

- We want to show that as time increases, the low frequency components of $\widehat{\nabla_{\tilde{H}^n} E_m}$ decrease to zero monotonically for a gradually increasing number of low frequency components. Note that $\|\widehat{\nabla_{\tilde{H}^n} E_m}\|_{\ell^\infty(\mathbb{Z})} \to 0$ as $t \to +\infty$. Because a common scale factor for $\widehat{\nabla_{\tilde{H}^n} E_m}$ does not have any effect on the geometry of the curve evolution, we scale $\widehat{\nabla_{\tilde{H}^n} E_m}$ by its maximum value in l to show the convergence to zero of low frequency components relative to the rest of the components of $\widehat{\nabla_{\tilde{H}^n} E_m}$. Consider

$$\tilde{g}_n(l,t) = \frac{g_n(l,t)}{1/(et)} = etg_n(l,t).$$

Note that

$$\frac{\partial \tilde{g}_n}{\partial t} = \frac{e}{\lambda(2\pi l)^{2n}} \exp\left(-\frac{t}{\lambda(2\pi l)^{2n}}\right)\left(1 - \frac{t}{\lambda(2\pi l)^{2n}}\right) < 0$$

when $l < \frac{1}{2\pi}(t/\lambda)^{1/(2n)}$. Note also that $\lim_{t\to+\infty} \tilde{g}_n(l,t) = 0$ for all $l \in \mathbb{Z}$. These facts verify our assertion.

- For a fixed time, we approximate the frequency component that is changing the most. To do this, we suppose l is a real number (even though it is an integer) and compute the derivative,

$$\frac{\partial g_n}{\partial l} = \frac{2}{\lambda(2\pi)^{2n} l^{2n+1}} \exp\left(-\frac{t}{\lambda(2\pi l)^{2n}}\right)\left[-1 + \frac{t}{l^{2n}}\right],$$

and so $l_* \approx \pm t^{\frac{1}{2n}}$ is the frequency component being changed the most at a particular time instant. In particular, this shows a successive transition through *all* possible frequencies in a coarse-to-fine manner (see the right plot in Fig. 20).

In the left of Fig. 20, we show a plot of $\tilde{g}_n$ to visually show the movement of $\widehat{\nabla_{\tilde{H}^n} E_m}$ from coarse to finer scale perturbations. The figure shows the case when $n = 2$; however, other $n > 0$ have the same qualitative behavior. Larger n shows a slower and more pronounced progression from coarse to fine. Note that when $n = 0$, $\tilde{g}_n(l,t) = \frac{et}{\lambda}\exp\left(-\frac{t}{\lambda}\right)$, which is constant in l for a fixed time, and so the plot on the left of Fig. 20,

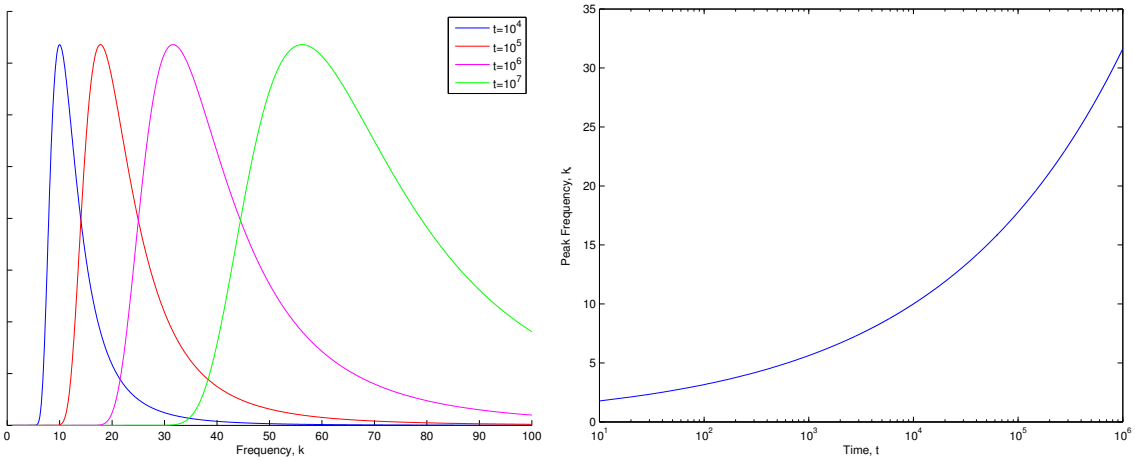


Figure 20: Left: Plot of $\tilde{g}_n(\cdot, t)$ with $n = 2$ for various t, which shows coarse-to-fine behavior. Right: Plot of frequency component changing most rapidly versus time, which shows that the evolution transitions through *all* possible frequencies.

would simply be a horizontal line for all time. The right of Fig. 20 shows (when $n = 2$) the frequency for which $\tilde{g}_n(\cdot, t)$ (or $g_n(\cdot, t)$) is maximized versus time. This plot shows that the Sobolev active contour transitions through all possible frequencies.

The evolutions (H^0 and $\tilde{H}^2$) minimizing E_m are shown in Fig. 21. Notice that the H^0-evolution favors all frequency components the same, and so the coarse and fine features of the target curve are both detected at the same times. On the other hand, the Sobolev evolution changes in a coarse manner, detecting the coarse structure of the target curve, before finally detecting the fine "spiked" structure and deforming in a fine-scale manner.

Remark 4.3.2. *For real images, as long as the image data does not undergo drastic changes with respect to movements of the curve, this coarse-to-fine motion shown in this example also applies to real image-based evolutions (in particular, for many tracking applications). In the general case, we would expect this coarse-to-fine deformation to repeat during successive intervals of time.*

The next energy we consider is the following smoothing energy, $E_s : C^\infty(S^1, \mathbb{R}^2) \rightarrow$

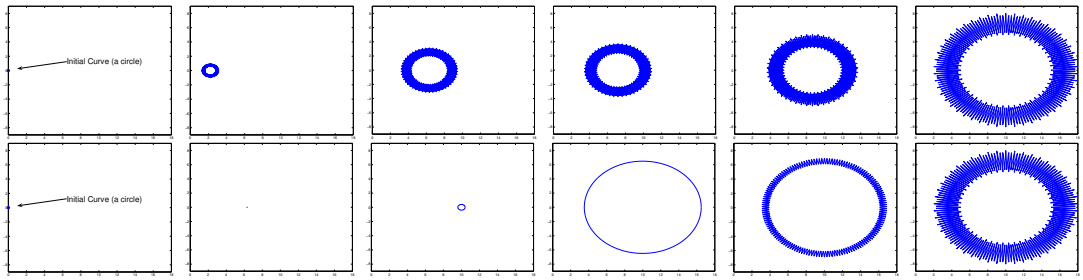


Figure 21: Minimizing E_m using a "spiked circle" (which is shown on the right of each row) for c_0, and the initialized curve is a circle: $c(u,0) = \epsilon(\cos 2\pi u, \sin 2\pi u), \epsilon = 0.001$ for $u \in S^1$ (left of each row, enlarged for visibility). Top: Snapshots of the H^0 evolution. Bottom: Snapshots of the $\tilde{H}^2$ evolution. The right of each row shows that both flows converge to the desired "spiked" circle, but taking quite different paths.

$\mathbb{R}^+$:

$$E_s(c) = \frac{1}{2}\int_{S^1} |c^{(m)}(u)|^2 \, \mathrm{d}u, \tag{99}$$

where $m \geq 1$.

Remark 4.3.3. *The geometric equivalent to* (99) *is*

$$E(c) = \frac{1}{2}\int_c |D_s^m c(s)|^2 \, \mathrm{d}s.$$

In the case that $m = 1$, the energy is half of arc-length, and in the case $m = 2$, the energy is the elastic energy. The Sobolev flows for these energies were calculated in Chapter 3, Section 3.4 . As we progress, we will illustrate the similar qualitative behavior between the parametric and geometric energies, which justifies our use of a parametric analysis to infer some qualitative behaviors in the geometric case.

Note that $\nabla_{H^0} E_s(c) = (-1)^m c^{(2m)}$; therefore,

$$-\widehat{\nabla_{H^0} E_s}(l) = -(-1)^m (2\pi i l)^{2m} \widehat{c} = -(2\pi k)^{2m} \widehat{c}(k)$$

and also,

$$-\widehat{\nabla_{\tilde{H}^n} E_s}(l) = \begin{cases} 0 & \text{for } l = 0 \\ -\lambda^{-1}(2\pi l)^{2(m-n)} \widehat{c}(l) & \text{for } l \neq 0 \end{cases}.$$

We would like to compute the gradients as a function of time in Fourier domain. Note, $\partial_t \widehat{c}(l,t) = -\widehat{\nabla_{\tilde{H}^n} E_s}(l)$, so

$$\partial_t \widehat{c}(l,t) = \begin{cases} 0 & l = 0 \\ -\lambda^{-1}(2\pi l)^{2(m-n)} \widehat{c}(l) & l \neq 0 \end{cases},$$

which yields a solution of

$$\widehat{c}(l,t) = \begin{cases} \widehat{c}(0,0) & l = 0 \\ \exp\left(-\lambda^{-1}(2\pi l)^{2(m-n)} t\right) \widehat{c}(l,0) & l \neq 0 \end{cases}.$$

Hence, we see

$$\widehat{\nabla_{\tilde{H}^n} E_s}(l,t) = \begin{cases} 0 & l = 0 \\ \lambda^{-1}(2\pi l)^{2(m-n)} \exp\left(-\lambda^{-1}(2\pi l)^{2(m-n)} t\right) \widehat{c}(l,0) & l \neq 0 \end{cases}.$$

We make the following observations:

- For $n < m$, the rate of convergence of $\widehat{\nabla_{\tilde{H}^n} E_s}(l,\cdot)$ to 0 (also $\widehat{c}(l,\cdot)$ to 0) increases as $|l|$ increases. Using similar arguments as for E_m, one can see that this means that $\widehat{\nabla_{\tilde{H}^n} E_s}$ moves from fine to coarse scale perturbations. Note that this case includes the case when $m = 1$ and $n = 0$, which is linear heat flow. This flow is well-known to have a smoothing effect on the curve, and removes fine scale curve information before removing coarser scale information (i.e., $\widehat{\nabla_{\tilde{H}^n} E_s}$ moves from fine to coarse). Similarly, in the geometric case when $m = 1$ (i.e., the energy is the length of the curve) and $n = 0$, the evolution is non-linear geometric heat flow [40, 37], which is well-known to have a fine to coarse smoothing effect.

- For $n = m$, the rate of convergence of all frequency components of $\widehat{\nabla_{\tilde{H}^n} E_s}(l,\cdot)$ to 0 is the same for all l. The curve evolution is a simple rescaling of the contour about its parametric centroid. It is also apparent in this case that the curve evolution exists when the gradient *ascent* flow is considered. Similarly, as

verified in Chapter 3, Section 3.4 through direct computations, in the case of the geometric energy when $m = 1$ and $n = 1$, the evolution is a simple rescaling of the curve about the *geometric* centroid, and therefore also stable for the ascent.

- For $n > m$, the rate of convergence of $\widehat{\nabla_{\tilde{H}^n} E_s}(l, \cdot)$ to 0 (also $\widehat{c}(l, \cdot)$ to 0) decreases as $|l|$ increases. As in the case of the energy E_m, we can show that $\widehat{\nabla_{\tilde{H}^n} E_s}$ moves from coarse to finer scale motions.

- From the above statements, we see that the Sobolev gradient flows move in a more coarse to fine way than the H^0 gradient flow, and as the order of the Sobolev gradient increases, this coarse-to-fine motion is more pronounced.

4.4 Benefits of Sobolev Contours for Tracking

In this section, we outline the benefits of switching from the standard H^0 active contour evolution to a Sobolev active contour in tracking algorithms that use active contours.

The scale-space analysis of Sobolev active contours performed in Section 4.3 that shows a coarse-to-fine evolution of the contour also shows why Sobolev active contours are ideal for tracking. The fact that H^0 gradient flows change fine structure of the curve immediately when energetically favorable, and hence are easily attracted by undesirable local minima, is one reason for predicting motion and dynamics of the object being tracked. By predicting motion and dynamics of the moving object, a better estimate of the object's upcoming position can be attained thereby placing the initial guess hopefully closer to its desired final position. Many prediction schemes apply low dimensional global motions to the contour. Thus, the initial global motion followed by an H^0 flow is less likely than the naive tracker to get caught in an intermediate, undesirable local minimum of the energy. Notice that since Sobolev gradient flows naturally move from coarse to successively finer motions, the contour is less likely to be trapped by intermediate local minima caused by local features of

the image, and is therefore likely to be less dependent on the prediction of motion and dynamics of the object. We also wish to emphasize that the transition from coarse to increasingly finer motions is automatic and continuous in comparison to other works (e.g., [99]) where the global motions must be deliberately specified, and the transition from the global motion to more local deformation is not continuous. Indeed, even discrete attempts to deliberately graduate from more global to more local motions are not trivial as one typically starts from translations, then rotations, then scale, but beyond this it becomes less clear and natural how to progress to finer scale deformations.

Another advantage of using Sobolev active contours for tracking is speed of convergence compared to standard H^0 active contours. While computing the $\tilde{H}^n$ gradient is slightly more computationally costly than computing the H^0 gradient, though both have the same order of complexity, we point out that without accurate prediction, the number of iterations in typical contour tracking applications required to update the active contour from frame to frame is usually much smaller with Sobolev active contours. Therefore the total computational time for processing between frames is significantly lower with Sobolev active contours. The reason is that the frame-to-frame motion of the object to be tracked is, as mentioned previously, usually dominated by more global motions: translations, scaling, and coarse scale deformations. Accordingly, a Sobolev active contour needs only a few iterations to lock onto the object in the next frame because the Sobolev gradient moves globally at first, preferring coarse scale motions in the first few iterations before proceeding to fine scale motions in later iterations. In contrast, standard H^0 active contours requires many more iterations since they immediately deform by local motions, significantly changing their initial shape (often to meaningless intermediate shapes), before deforming back to only slightly deformed, translated and scaled versions of their initial shape, and that is assuming they don't first get trapped into intermediate local minima!

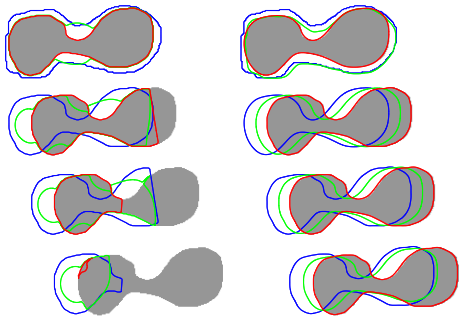


Figure 22: Simple tracking using geodesic active contours: Standard (H^0) active contour (left column) deforms the initialized contour greatly and is stuck in local minima, and Sobolev active contour (right column) moves in a global manner only slightly changing shape. In each frame, the initial curve (given by the contour detected in the previous frame) is blue, the intermediate curve is green, and the final detected curve is red.

We now illustrate the advantages discussed in the previous paragraphs with a simple synthetic image sequence (Fig. 22) in which we employ the naive tracker using the energy functional for geodesic active contours [16, 53]:

$$E_{geo}(c) = \int_c \phi(c(s))\,\mathrm{d}s, \text{ where } \phi = \frac{1}{1+\|\nabla I\|^2}. \tag{100}$$

Fig. 22 shows the tracking for both the H^0 gradient flow and the $\tilde{H}^1$ gradient flow. The flows are run until convergence in each frame. Note that the H^0 active contour deforms its initial shape greatly to react to local information. Hence the contour changes shape and must re-deform back to its initial shape. However, the contour gets trapped in an undesirable local minimum. The Sobolev active contour, on the other hand, only changes shape slightly while moving in an overall translation. This means that the number of iterations until convergence for the H^0 active contour is much greater than the Sobolev active contour, and therefore the computational time is also much greater. See Fig. 23 for a simple quantitative analysis of the number of iterations and computational times. In this simulation, we segment the object shown in Fig. 22 when the initial contour is a translated and a slightly deformed version of the object. We quantify the difference by using the set symmetric difference between the desired object and the initial contour. From the graph in Fig. 23, we see that

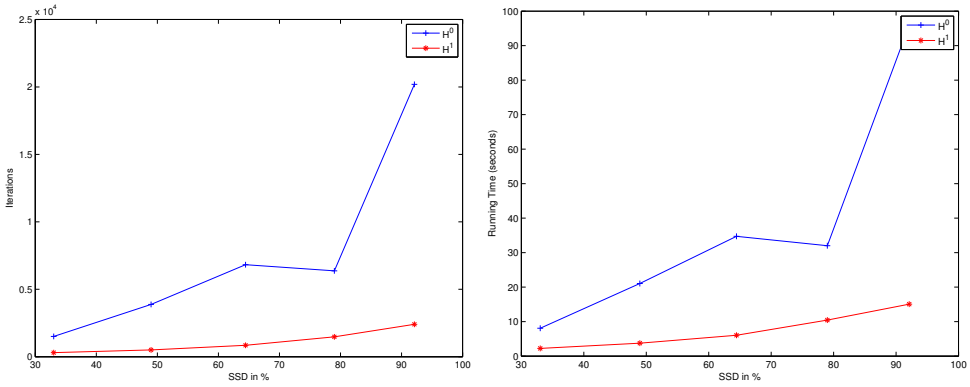


Figure 23: Graphs showing number of iterations to converge and total time for convergence versus set symmetric difference (SSD) of initial region and desired object in percent (scaled by desired object area).

the number of iterations and the computational time is significantly lower for the $\tilde{H}^1$ active contour.

4.5 Experiments

We now demonstrate significant performance gains by replacing standard H^0 active contours with their Sobolev counterparts for the *exact same* detection energy in a variety of tracking scenarios on real videos, both when using the *naive tracker* as well as when tracking with a predictor. These experiments give evidence to support our claim that the Sobolev metric rather than the traditional metric should be used in tracking applications that make use of active contours.

Note that in the next experiments, we use the $\tilde{H}^1$ active contour and the algorithm described in Chapter 3, Section 3.3.4 that is independent of the parameter λ in the definition of the $\tilde{H}^1$ inner product. The algorithm evolves by the translation component of the $\tilde{H}^1$ gradient until this term becomes zero followed by one iteration of the deformation component, which is geometrically independent of λ, and the process is iterated.

Fig. 24 shows the results for a sequence in which a man is walking on a street. The sequence is heavily corrupted by noise (Gaussian noise: $\mu = 0, \sigma^2 = 0.3$). The

tracking is done using the naive tracker (no prediction) where the detection energy is the Chan-Vese energy functional [17]:

$$E_{cv}(c) = \int_{c_{in}} (I-u)^2 \, \mathrm{d}A + \int_{c_{out}} (I-v)^2 \, \mathrm{d}A + \alpha L(c), \tag{101}$$

where

$$u = \frac{\int_{c_{in}} I \, \mathrm{d}A}{\int_{c_{in}} \mathrm{d}A}, \text{ and } v = \frac{\int_{c_{out}} I \, \mathrm{d}A}{\int_{c_{out}} \mathrm{d}A},$$

and $\alpha \geq 0$ specifies a penalty on the length (used for curve regularity for the H^0 active contour) $L(\cdot)$ of the curve. The top row shows the standard H^0 ($\alpha = 5000$) active contour, and the bottom row shows the Sobolev $\tilde{H}^1$ ($\alpha = 0$) active contour. Sobolev active contours in particular favor translations; therefore in order to show that the translation favoring property of Sobolev active contours is not solely responsible for the pleasing tracking results, but more generally it is the coarse-to-fine property, we also show the results of tracking where the energy minimization is performed using an H^0 inner product that has a heavily weighted translation component in the middle of Fig. 24 (note the advantages of the Sobolev technique over such explicitly favored groups was discussed at the end of Section 4.1.1); a similar result is obtained for an affine favored H^0 gradient. We have used an alternating algorithm between a translation and the H^0 gradient minus the translation, to avoid picking the weight on the translation, but the result is similar to Fig. 24 (middle row). Note that such an inner product was considered in [21]. The contours are evolved until convergence between frames. After a few frames, the H^0 active contour gets stuck in noise and loses track of the person. The translation favored H^0 contour initially does better than the H^0 active contour, but soon loses track of the person, becoming stuck in noise. Note that the translation, in the first frames, initially pushes the contour in the vicinity of the person, but then the active contour immediately detects the fine scale noise since H^0 minus the translation does not favor coarser motions, and cannot more accurately detect the person. The Sobolev active contour, because of

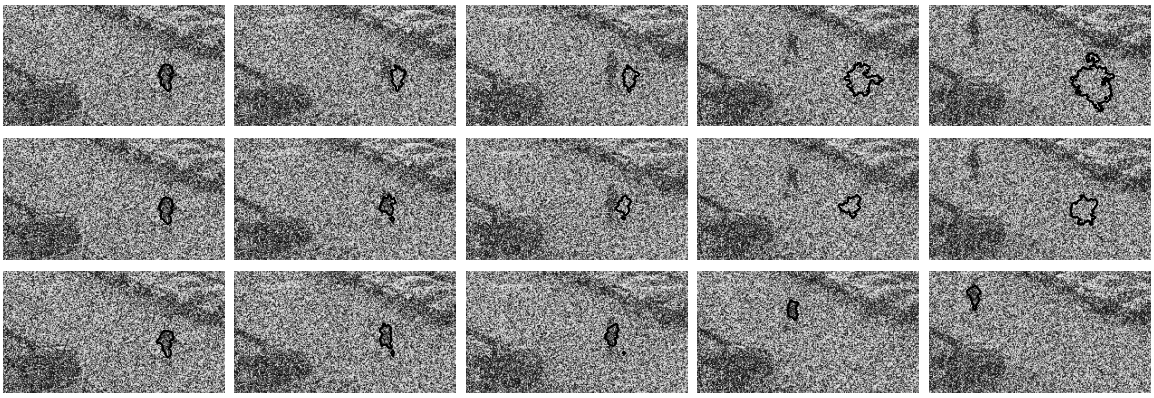


Figure 24: Tracking of a person in a noisy image sequence with a region-based (Chan-Vese) energy with H^0 (top row), with H^0 translation favored (middle row), and with $\tilde{H}^1$ (bottom row) active contours.

its more global initial motions (translations and other coarse motions), avoids the intermediate local minima caused by noise and keeps tracking the person. Due to the high noise level, however, the precise shape of the person is not captured in all of the cases. Note that one may try different weights, α, to obtain better results for the H^0 and the translation favored H^0 active contours; however, we have found that lower values (than $\alpha = 5000$) produce results that are more easily trapped in noise and higher values simply shrink the curve to a point. It should be noted that although a length penalty for the Sobolev $\tilde{H}^1$ is not needed (and indeed does not make the curve more regular: see Chapter 3, Section 3.4), a length penalty of $\alpha = 5000$, does not significantly affect the results shown. Of course a much higher α would also shrink the curve to a point.

To quantify the robustness to noise of the Sobolev active contour versus the traditional H^0 active contour as seen in the previous experiment, we have conducted experiments with a synthetically generated image sequence (so that the ground truth is known) in which various degrees of noise are added. The sequence is binary images in which a square is translating (to represent motion) and changing its area slightly (to represent deformation). In the first experiment, we use the naive tracker and the detection energy (100). In this case the square is translated by three pixels and the

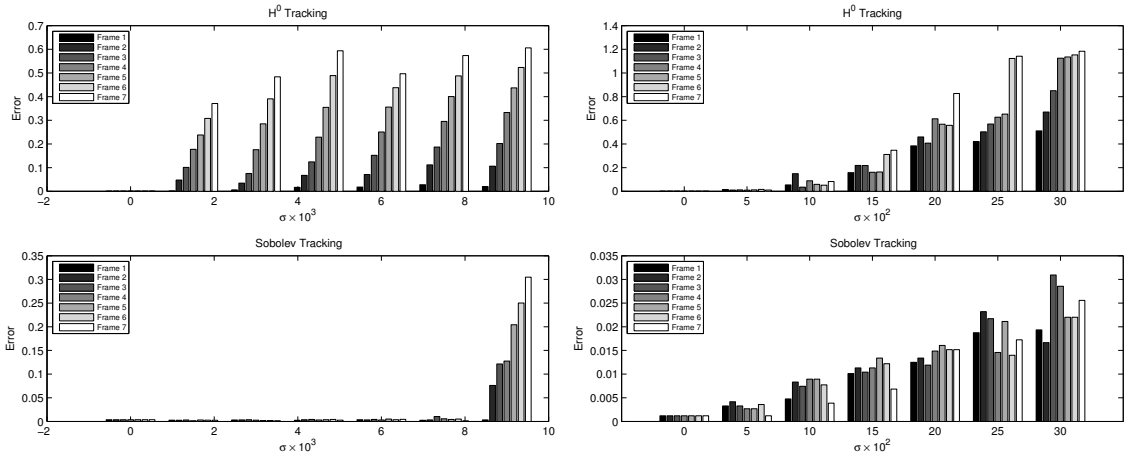


Figure 25: Plots of error for tracking a square that is translating and slightly changing its area with various degrees of noise (Gaussian mean 0, standard deviation σ). Left: Using geodesic active contours, right: using the Chan-Vese model. Note the difference in the scales of each plot; in particular, the plots show that the results of using Sobolev active contours is vastly better than the corresponding H^0 active contour.

length of the side is randomly changed by ± 2 pixels when compared with the square in the previous frame. The segmentation error for various degrees of Gaussian noise ($\mu = 0$, and standard deviation σ specified) using both H^0 and Sobolev $\tilde{H}^1$ active contours is shown in the left of Fig. 25. Similar experiments are done using the energy (101), but the square is translated by 17 pixels and the length is adjusted by a random ± 5 pixels. Results are shown in the right of Fig. 25. The results for the H^0 active contour are shown for the best value of α chosen for the given noise levels. Note that our measure of error is one-half of the number of false positive classified pixels plus false negative classified pixels divided by the ground truth number of pixels of the object. In both cases of the detection energies chosen, the Sobolev active contour does significantly better than the corresponding H^0 active contour.

In the next experiment (Fig. 26), we demonstrate that the Sobolev active contour is useful not only for noisy situations, but in other cases where one is trying to track an object in a cluttered or textured environment where the object shares some visual characteristics with the background. In this experiment, we track a sea creature at

the bottom of an ocean using the naive tracker and the detection energy,

$$E_{cv+\sigma}(c) = (1-\beta)E_{cv}(c) - \frac{1}{2}\beta(\sigma_u^2 - \sigma_v^2)^2$$

where

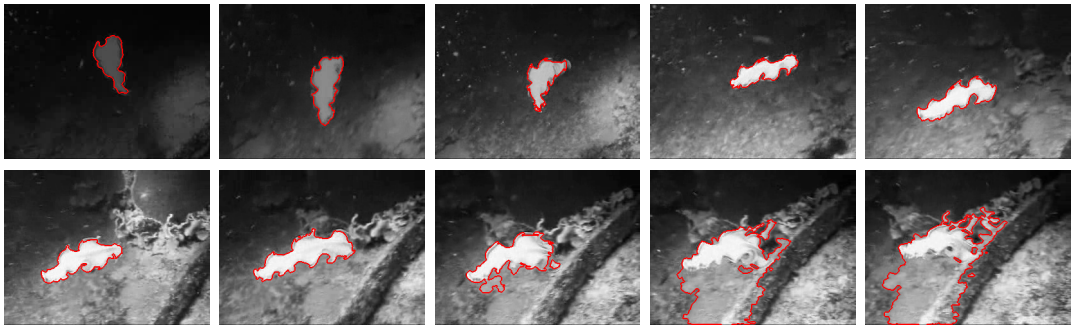
$$\sigma_u^2 = \frac{\int_{c_{in}} (I-u)^2 \, \mathrm{d}A}{\int_{c_{in}} \mathrm{d}A} \text{ and } \sigma_v^2 = \frac{\int_{c_{out}} (I-v)^2 \, \mathrm{d}A}{\int_{c_{out}} \mathrm{d}A}.$$

Since the mean values of some regions in the background are closer to the mean value inside the creature rather than to other regions of the background, a first order Chan-Vese energy is not enough to capture the object, and thus, we incorporate second order information. For this experiment, we chose $\beta = 0.6$ although different β produced similar results. In Fig. 26, we see that the H^0 active contour tracks the object for some time, but when the object's statistics look closer to the light part of the background than the dark part, the contour leaks into the background. On the other hand, because the Sobolev active contour moves globally before gradually changing its fine structure, the contour is able to avoid the distracting features of the background and get a rough approximation of the object before detecting finer features of the object, and thereby locking into a more desirable local minimum than the H^0-active contour.

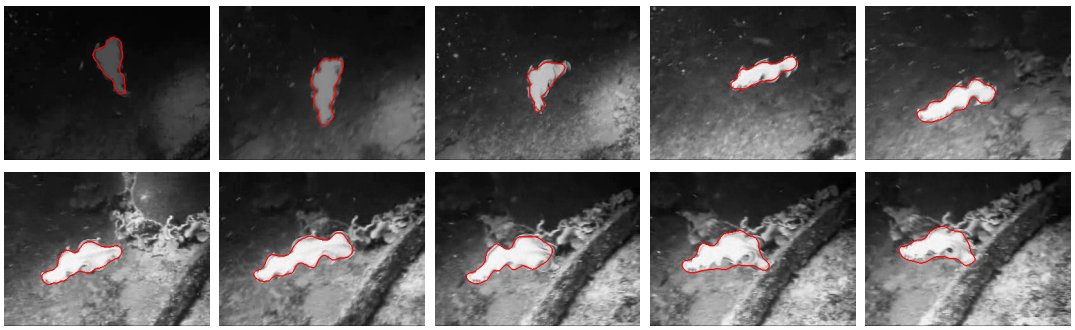
In the next experiment (in Fig. 27), we show that the Sobolev active contour can offer improvements over the traditional metric for tracking an object through partial occlusions. In particular, we track a car that moves under a lamp post. The energy functional used for the active contours is the Mumford-Shah functional [72]:

$$E_{ms}(c, f, g) = \int_{c_{in}} (I-f)^2 \, \mathrm{d}A + \int_{c_{out}} (I-g)^2 \, \mathrm{d}A \tag{102}$$

where f and g are smooth functions defined inside (resp. outside) the curve. The functional is minimized jointly in c, f, and g (see [106, 28, 110] for implementation details). A fixed number of iterations (300) is used to evolve the curve at each frame. The top row shows the H^0 active contour, which is thrown off as soon as the contour



(a) Tracking using the H^0 active contour.



(b) Tracking using the Sobolev ($\tilde{H}^1$) active contour.

Figure 26: Tracking of a sea-creature at the sea bottom using an energy which incorporates the mean intensity and variance information inside and outside the contour.

Figure 27: Tracking of a car under an occlusion using the Mumford-Shah energy with H^0 (top) and $\tilde{H}^1$ active contours.

hits the lamp post. This is because each point of the H^0 active contour moves in a direction independent from the other points. Hence, the points close to the lamp post do not want to move past the post. On the other hand, the Sobolev $\tilde{H}^1$ flow moves globally first, and hence does not get stuck on the lamp post and continues to track the car, although at the end, the contour misses the outer parts of the car.

In the last experiment, we illustrate that the Sobolev active contour can improve the traditional active contour even when a prediction step is used to obtain the overall global motion (indeed an affine motion is predicted). The experiment (in Fig. 28) tries to address the problem with the previous experiment by using a predictor and observer/estimator. We use the detection/prediction algorithm considered in [49]. The detection step involves a simultaneous segmentation and rigid registration (i.e., affine) of three consecutive frames using the Mumford-Shah functional (102) and a fixed (300) number of iterations. For the prediction, a constant acceleration model is assumed for the parameters of the rigid registration. The measurements that the estimator uses to estimate the contour and its registrations are just the results of the detection step. A Kalman gain is used to determine if more weight is put on the measured contour versus the model prediction. As can be seen in Fig. 28, the H^0 active contour prefers coarse scale and fine scale perturbations equally and therefore the contour immediately becomes distracted by the pole, which is detected by fine scale perturbations, and the estimator/predictor is of little help. Note that a higher

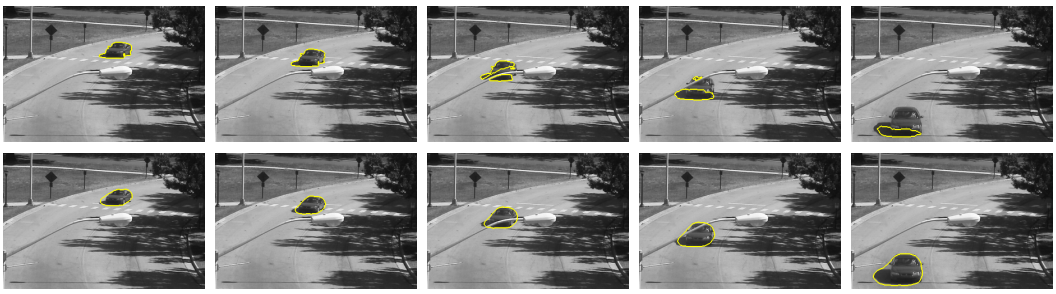


Figure 28: Tracking a car under an occlusion using estimation with Mumford-Shah energy functional for the detection. H^0 (top) and $\tilde{H}^1$ (bottom) active contours.

regularization penalty may be used; however, like the previous experiments, the length penalty, in addition to restricting the deformation into the pole, also shrinks the curve and a significant portion of the car is not detected. On the other hand, for the Sobolev $\tilde{H}^1$, the estimator/predictor greatly improves the result, as the shape is more accurately captured. Note that if the detection step is iterated for a very large number of iterations, the Sobolev active contour will also be distracted by the pole in this experiment. However, in real-time tracking applications it is often the case that the detection is not run until convergence; therefore, it is nice to know that with a limited number of iterations, the Sobolev active contour detects the coarse deformations, which are more essential than the fine deformations. In both cases (with and without the predictor/estimator), it is clear that simply replacing the standard H^0 active contour with the Sobolev active contour greatly improves the tracking performance.

4.6 Conclusion

We have shown that Sobolev active contours move successively from coarse to fine scale motions in a continuous manner through a scale-space type analysis. This property gives more justification for using the Sobolev framework. We have shown that this property, along with others, makes Sobolev active contours natural for tracking, and experiments have shown that the Sobolev technique is beneficial over the

standard technique both when tracking with or without a predictor. The property of coarse to fine motion, as we saw, implies that Sobolev active contours take fewer iterations (and also less time) to converge to the desired local minimum than H^0 active contours. This is important for real-time tracking systems where a more efficient detection scheme with better accuracy is beneficial. Note that existing tracking algorithms, which use active contours, need not be modified; nor does the energy functional for the active contour, just a simple addition of a procedure to compute the Sobolev active contours is necessary, which is straight forward to obtain from the original active contour.

In this article, we have analyzed Sobolev gradient flows for curves and showed important properties of these flows, which are quite useful for tracking applications. The next step is to extend these ideas to surface evolutions, where there are many applications such as tracking of the heart in ultrasound sequences. Although the Sobolev method extends to surfaces, there are no equivalent convolution formulas (and in particular no simple integral solution as in $\tilde{H}^n$) for Sobolev gradients with respect to surfaces, and they are computationally expensive to compute. Future work is to formulate computationally feasible methods for computing the Sobolev gradients for surfaces, or consider alternate definitions of the Sobolev metric that lead to fast computations of the corresponding gradient.

CHAPTER V

FURTHER APPLICATIONS OF SOBOLEV ACTIVE CONTOURS

Some interesting properties of Sobolev gradient flows are that they stabilize certain unstable traditional flows, and the order of the evolution PDEs are reduced when compared with traditional gradient flows of the same energies. In this chapter, we explore new possibilities for active contours made possible by Sobolev active contours. The Sobolev method allows one to implement new energy-based active contour models that were not otherwise considered because the traditional minimizing method cannot be used. In particular, we exploit the stabilizing and the order reducing properties of Sobolev gradients to implement the gradient descent of these new energies. We give examples of this class of energies, which include some simple geometric priors and new edge-based energies. We will show that these energies can be quite useful for segmentation and tracking. We will show that the gradient flows using the traditional metric are either ill-posed or numerically difficult to implement, and then show that the flows can be implemented in a stable and numerically feasible manner using the Sobolev gradient.

5.1 Introduction

The main purpose of Chapter 3 and Chapter 4 was to show advantages of using Sobolev active contours over the traditional active contour based on the same energy. In contrast, in this chapter we introduce *new* active contour energies that are quite useful for various segmentation tasks, but *cannot* be minimized with the traditional H^0 active contour (nor other global optimization techniques), and the Sobolev active

contour must be used. We show a few examples of these energies, which include simple geometric priors for active contours and new edge-based energies. These new energies fall into two categories: one in which the resulting H^0 flows are not stable, and another in which the traditional gradient flow results in high order PDEs that are numerically difficult to implement using level set or particle based methods. We propose to use Sobolev active contours, which avoid both of these problems.

This chapter is meant to illustrate that energies that result in H^0 unstable or high order flows can still be considered for optimization with the Sobolev method (and these energies need not be discarded or adjusted). As such we give a few simple examples of the energies that fall into these categories. Experiments in this chapter show the types of behaviors that can be obtained from the simple energies considered, and one can obtain good results on more complex images by combining these results with other energies.

The graph cut method, a global minimization technique for minimizing discrete approximations of some active contour energies, is known to be able to minimize three types of geometric energies: weighted length, flux of a vector field, and weighted area [56]. Some of the energies we consider are non-simple operations (such as division) of the previous energies mentioned, and the technique in [56] does not apply; indeed we are not certain that a graph for the energies we consider can be constructed. Other energies we consider have curvature inside the integral; since edge weights in graph cuts depend on an edge (between a pixel and its neighbor), it is unclear whether a curvature term may be incorporated in the framework since to compute curvature one needs three points. In any case, since graph cut methods do not have sub pixel accuracy, curvature computations would be extremely inaccurate.

In [19], the authors consider various different metrics resulting in 'coherent' gradient flows; indeed they construct flows that favor certain group motions such as affine motions. In the case of the affine group (others are analogous), the flow is formed by

re-weighting the affine component of the traditional gradient higher and the component orthogonal (according to the H^0 inner product) lower. For the class of energies that we wish to explore in this chapter however, the metrics proposed by [19] based on group motions also suffer from the same problems as the traditional H^0 metric; namely, these flows are either not stable or are high order PDEs and are difficult to implement numerically.

5.2 Some Useful Energies Precluded by H^0

In this section, we introduce three geometric "energies", which can be used as building blocks to produce a variety of other useful energies (to be described in subsequent sections). We then derive the H^0 gradient and show that the gradient descent flow is either ill-posed or very difficult to implement numerically. We then derive the Sobolev gradient flows, and justify that they are well-posed and numerically feasible to implement.

Before we proceed, we introduce the notation used in this chapter. A contour will be denoted, c; its length is L. We define $\bar{\cdot} := L^{-1}\int_c \cdot \,\mathrm{d}s$, where $\mathrm{d}s$ is the arc length measure of c. Throughout this chapter, we define the Sobolev metric as $\|h\|^2_{\text{Sobolev}} := |\text{avg}(h)|^2 + \lambda\|h'\|^2_{H^0}$, where h is a perturbation of c, h' is the arc parameter derivative, and $\lambda > 0$ is a scaling factor. We note the fact from Chapter 3, Section 3.3.1 that the Sobolev gradient can be computed from the H^0 gradient as $\nabla_{\tilde{H}^1} E = K * \nabla_{H^0} E$, where $*$ is circular convolution and K is found in Chapter 3, Section 3.3.1 whose second derivative exists in a distributional sense.

The first "energy" that we introduce is the following generalization of average weighted length:

$$E(c) = \frac{1}{L}\int_c \phi(c(s))\,\mathrm{d}s = \text{avg}(\phi), \tag{103}$$

where $\phi : \mathbb{R}^2 \to \mathbb{R}^k$ where $k \geq 1$. The H^0 gradient of this energy is

$$\nabla_{H^0} E(c) = \mathcal{N}[\mathcal{N}^T (D\phi)^T - \kappa(\phi - \text{avg}(\phi))^T] \tag{104}$$

where T denotes transpose, and D denotes derivative. Since $\phi - \text{avg}(\phi)$ is not strictly positive, the gradient descent flow has a component that is reverse heat flow on half of the contour, and therefore the H^0 gradient descent is ill-posed. Note that the reverse heat component attempts to increase the length of certain portions of the contour. Since the ill-posedness of the H^0 flow only arises from the length increasing effect, we expect the Sobolev gradient flow to be well-posed. This is because increasing the length of the contour is a well-posed process using the Sobolev gradient; indeed, the Sobolev gradient ascent for length is simply a rescaling of the contour (see Chapter 3, Section 3.4). Computing the Sobolev gradient of (103) we have

$$\nabla_{\tilde{H}^1} E(c) = -\frac{c - \text{avg}(c)}{\lambda L^2}\text{avg}(\phi)^T + K * (D\phi)^T + K' * (c_s \phi^T) \tag{105}$$

Notice that the component, $\mathcal{N}\text{avg}(\phi)^T\kappa$, of the H^0 gradient that caused the ill-posedness has been converted to the first term of the Sobolev gradient (105), which is a stable rescaling of the contour.

Next, we introduce a scaled version of the weighted area, given by the energy

$$E(c) = \frac{1}{L^2}\int_R \phi(x)\,\mathrm{d}A(x) = \frac{A_\phi}{L^2}, \tag{106}$$

where $\phi : \mathbb{R}^2 \to \mathbb{R}$, R is the region enclosed by c, and $\mathrm{d}A$ is the area measure in $\mathbb{R}^2$. Similar to the previous energy, the ill-posedness of the H^0 gradient descent flow of (106) is due to the scale factor of L^{-2}, which causes a length increasing component in the gradients, and is ill-posed with respect to H^0. Indeed, calculating the gradient, we have

$$\nabla E(c) = \frac{L^2 \nabla A_\phi - 2A_\phi L \nabla L}{L^4} = \frac{A_\phi}{L^2}\left[\frac{\nabla A_\phi}{A_\phi} - 2\frac{\nabla L}{L}\right].$$

Therefore, we see that

$$\nabla_{\tilde{H}^1} E(c) = -\frac{A_\phi}{L^2}\left[L\frac{K * (\phi\mathcal{N})}{A_\phi} + 2\frac{c - \text{avg}(c)}{\lambda L^2}\right], \tag{107}$$

which leads to a well-posed descent (and ascent).

Lastly, we introduce the following generalization of the elastic energy:

$$E(c) = L \int_c \phi(c(s))\kappa^2(s)\,\mathrm{d}s, \tag{108}$$

where $\phi : \mathbb{R}^2 \to \mathbb{R}$, and κ is the signed curvature of c. The factor of L multiplying the integral makes the energy scale-invariant when ϕ is a constant. Note that without the L factor, one can make the elastic energy arbitrarily small by scaling a contour large enough. We will also consider the scale-varying elastic energy without the L. These energies have been used in the past for the "curve completion" problem, which is a curve interpolation problem between two points [47, 11]. In [11], for the numerical implementation, a discrete version of the energy is minimized with a "shooting" method. One can show that the H^0 gradient of (108) is

$$\nabla_{H^0} E(c) = -Ec_{ss} + 2L^2\partial_{ss}(\phi c_{ss}) + 3L^2\partial_s(\phi\kappa^2 c_s) + L^2\kappa^2\nabla\phi \tag{109}$$

We note the result of [83], which considers the H^0 gradient descent flow of an energy similar to (108). The author of [83] considers the H^0 gradient descent flow of the energy

$$E(c) = \int_c (\kappa^2(s) + \alpha)\,\mathrm{d}s$$

where $\alpha > 0$. It is proven that an immersed/regular curve evolving under this fourth-order flow stays immersed/regular, and a solution exists for all time. In the case when ϕ is a constant, the flow (109) is similar to the flow that is considered in [83], except that α is time varying in (109). For numerical implementation, the fourth order flow (109) is difficult to implement with marker particle methods because of numerical artifacts arising from fourth order differences, and it is even more problematic to implement with level set methods because the flow is not known to have a maximum principle and because of numerical artifacts. These are the reasons for considering the Sobolev gradient:

$$\nabla_{\tilde{H}^1} E = -\frac{E}{\lambda L^2}(\mathrm{avg}(c) - c) + \frac{2}{\lambda}(\mathrm{avg}((\phi\kappa\mathcal{N})) - \phi\kappa\mathcal{N}) - 3L^2K' * (\phi\kappa^2\mathcal{T}) + L^2K * (\kappa^2\nabla\phi). \tag{110}$$

The Sobolev flow is second order, although it is a integral PDE. We can bypass the question about a maximum principle for this flow since the local terms have a maximum principle, and we perform extensions in the level set implementation for global terms.

5.3 Geometric Priors for Active Contours

In this section, we introduce some simple geometric shape priors for use in active contour segmentation. As these energies are formed from the energies presented in the previous section, they cannot be minimized with the usual H^0 gradient descent.

5.3.1 Length and Smoothness Priors

In many active contour models, a curvature term, *i.e.*, $\alpha\kappa\mathcal{N}$ (where $\alpha > 0$ is a weight), is added to a data-based curve evolution. The resulting flow will inherit regularizing properties such as smoothing the curve from the addition of this term. If the active contour model is based on minimizing an energy, then adding a curvature term is equivalent to adding a length penalty to the original energy, that is, if E_{data} is the original energy then the new energy being optimized (w.r.t the traditional H^0 metric) is

$$E(c) = E_{\text{data}}(c) + \alpha L(c). \tag{111}$$

This may be considered as a simple prior in which we assume that the length of the curve is to be shrunk. In general segmentation situations, this assumption may not be applicable. A more general energy incorporating length information, when such prior length information is known, is

$$E(c) = E_{\text{data}}(c) + \alpha(L(c) - L_0)^2, \tag{112}$$

in which it is assumed that length of the target curve is near L_0. See [93] (and references within) for related flows where the length of the curve is preserved. Note

that this prior allows for increasing or decreasing the length of the curve based on the current length of the curve and the value of L_0. The H^0 gradient is

$$\nabla_{H^0} E(c) = \nabla_{H^0} E_{\text{data}}(c) - 2\alpha(L - L_0)\kappa\mathcal{N},$$

which leads to an unstable flow if $L - L_0 < 0$. The Sobolev gradient is

$$\nabla_{\tilde{H}^1} E(c) = \nabla_{\tilde{H}^1} E_{\text{data}}(c) + 2\alpha(L - L_0)\frac{c - \text{avg}(c)}{\lambda L},$$

which is stable if the data term is stable.

In active contour works, the goal of adding the length penalty may have been mainly for obtaining the regularizing properties of the resulting flow, even though the energy itself does not favor more regular curves. It is evident that the Sobolev length descent does not regularize the active contour since the flow is a rescaling of the curve. Thus, to introduce smoothness into the Sobolev active contour (and even the H^0 active contour), we introduce the smoothness prior given by the energy,

$$E(c) = E_{\text{data}}(c) + \alpha L(c) \int_c \kappa^2(s)\, \mathrm{d}s. \tag{113}$$

The energy itself favors smoother contours, and we are not relying on the properties of a particular metric for regularity; it is inherent in the energy itself. The factor of L is for scale-invariance (unlike the length descent, this regularizer does not favor shrinking the length of the contour). Using the scale-varying and scale-invariant elastic energies as smoothness measures for active contours is mentioned but not implemented in [29, 9].

5.3.2 Centroid and Isoperimetric Priors

We now consider incorporating prior information on the centroid, length, and area of a curve into active contour segmentation. We consider the energy

$$E(c) = E_{\text{data}}(c) + \alpha\|\text{avg}(c) - v\|^2 + \beta(L - L_0)^2 + \gamma(A - A_0)^2 \tag{114}$$

where $\alpha, \beta, \gamma \geq 0$ are weights, $\text{avg}(c)$ is the centroid of the curve c, $v \in \mathbb{R}^2$ is the centroid known *a-priori* (see Section 5.5.2 for an example of how this may be obtained), L_0 and A_0 are the prior values for the length and area. If detailed information is not known about the length and area, then that part of the energy may be replaced by the energy

$$E(c) = E_{\text{data}}(c) + \alpha \|\text{avg}(c) - v\|^2 + \beta(\rho(c) - \rho_0)^2 \tag{115}$$

where

$$\rho(c) = \frac{A(c)}{L^2(c)} \tag{116}$$

is the isoperimetric ratio, which is a geometric measure of the relative relation between the length and area of a curve. Note that ρ is scale-invariant. It is a well known fact that the isoperimetric ratio is maximized by circles, and the maximum ratio is $1/(4\pi)$. Thus, the prior ratio must be constrained so that $\rho_0 \leq 1/(4\pi)$. Note that a low (near zero) isoperimetric ratio can be obtained by a snake-like shape, and a high ratio implies a shape that looks close to a circle. The isoperimetric ratio is mentioned to be used as a smoothness measure in [29], but this idea is not pursued.

Note that both the H^0 gradient descents for the centroid constraint and the isoperimetric penalties are ill-posed. The isoperimetric ratio is a special case of (106) (when $\phi = 1$), and the constraint gives a gradient of $(\rho - \rho_0)\nabla\rho$, which gives an unstable H^0 gradient descent flow when $\rho > \rho_0$. Note that the centroid is a special case of (103) (when $\phi : \mathbb{R}^2 \to \mathbb{R}^2$ is $\phi(x) = x$). The gradient of the centroid penalty is $\nabla(\text{avg}(c))(\text{avg}(c) - v)$, which gives an H^0 gradient of

$$[(\text{avg}(c) - v) \cdot \mathcal{N} - (c - \text{avg}(c)) \cdot (\text{avg}(c) - v)\kappa]\mathcal{N}$$

using (104). The gradient descent is unstable when $(c - \text{avg}(c)) \cdot (\text{avg}(c) - v) < 0$. The Sobolev gradient using (105) is

$$(\bar{c} - v) + K' * [c_s(c - \text{avg}(c)) \cdot (\text{avg}(c) - v)].$$

One possible use for (114) and (115) is in tracking applications (see Section 5.5.2).

5.4 New Edge-Based Active Contour Models

The energy for the traditional edge-based technique [16, 53] (called geodesic active contours) is

$$E(c) = \int_c \phi(c(s))\,\mathrm{d}s \tag{117}$$

where $\phi : \mathbb{R}^2 \to \mathbb{R}$ is chosen low near edges (a common example is $\phi = 1/(1 + \|\nabla(G * I)\|)$ where G is a Gaussian smoothing filter). There are several undesirable features of this model (even if a perfect edge-map ϕ is chosen). The energy has trivial (undesirable) minima and even minima that are not at the edges of the image (see for example [62]). This is in part due to the bias that the model has in preferring shorter length contours, which may not always be beneficial. Therefore, we propose new edge-based models.

5.4.1 Non-Shrinking Edge-Based Model

We propose to minimize the following non-length shrinking edge-based energy:

$$E(c) = \int_c \phi(c(s)) \left(L^{-1} + \alpha L \kappa^2(s)\right)\,\mathrm{d}s, \tag{118}$$

where $\alpha \geq 0$, which we claim alleviates some of the undesirable properties of (117). An energy, which is similar to (118) (except for the factor of L on the curvature term), is considered by [34], but a discrete version of the energy is used for implementation. The first term, $\frac{1}{L}\int_c \phi\,\mathrm{d}s$ (*i.e.*, (118) when $\alpha = 0$), is the same as the energy used for the geodesic active contour model, but there is a scale factor of $1/L$. This removes the length shrinking effect of (117) in descent flows; in particular if there are no edges (ϕ is constant), then a descent flow will not shrink the contour. The H^0 gradient of the first term (when $\alpha = 0$ in (118)) as noted in (104) is $-L(\phi - \mathrm{avg}(\phi))\kappa\mathcal{N} + L(\nabla\phi\cdot\mathcal{N})\mathcal{N}$, which is zero when the contour is aligned on true edges of the image (note that this may not be the case with the geodesic active contour model). The flow is stable with respect to the Sobolev metric, but not with respect to H^0.

Dividing the energy (117) by L, as in the first term of (118), loses regularizing effects of the original flow, and it is possible that the contour can become non-smooth from irrelevant noise. This observation is the reason for the second term of (118). The second term, $L\int_c \phi\kappa^2\,\mathrm{d}s$, is an image dependent version of the scale-invariant elastic energy. This term favors smooth contours, but smoothness is relaxed in the presence of edges, which are determined by ϕ. The factor of L makes the energy scale-invariant when ϕ is constant; therefore, a descent flow will not increase or decrease the length of the contour unless these behaviors make the curvature smaller or make the contour align along the edges. The reason for not considering this term alone is for the following reason. Suppose we are considering open curves with two endpoints fixed. Regardless of the ϕ that is chosen, the minimum of this term is always zero, and it is minimized by a straight line (the curvature is zero). For closed contours, we have observed in the numerical implementation that the contour sticks to isolated points where there is an edge of the image, and the converged contour is a straight line between these points (even if there is no edge along the line). Thus, the contour looks polygon-like. Even though the $\kappa = +\infty$ at vertices of polygons, this is not true numerically where κ is finite. Therefore, in a numerical implementation, the second term of (118) is not useful by itself.

5.4.2 Increasing Weighted Length

Instead of a non-shrinking edge-based model, if we have prior information that the length of the curve should increase, *e.g.*, the initial curve is within the object of interest, then one may want to *maximize* the following energy:

$$E(c) = \int_c \phi(c(s))\,\mathrm{d}s - \alpha\int_c \kappa^2(s)\,\mathrm{d}s \tag{119}$$

where $\alpha \geq 0$, and ϕ, contrary to the geodesic active contour model, is designed to be large near edges (one example is choosing $\phi = \|\nabla I\|$). The first term of the energy is weighted length, and therefore this term favors increasing the length of the

curve while stopping near edges. Considering only the first term ((119) when $\alpha = 0$), since the length of the curve is being increased, it is likely that when the curve has converged on a coarse scale, fine details due to noise become detected and the curve becomes rough, thereby further increasing length. Therefore, we add a regularizer, which is the second term of (119), to the weighted length. Note that we propose to use the scale-varying elastic energy, which in addition to regularity, gives an effect of increasing the length of curve, which is beneficial based on the prior assumption.

The H^0 gradient ascent of the weighted length term results in one term that is $-\phi\kappa\mathcal{N}$, which makes the length of the curve increase and is unstable. If $\alpha > 0$, then the H^0 flow of (119) may become well-posed since this results in higher order regularity terms, but the elastic energy has its own problems using the H^0 gradient flow. Therefore, we use the Sobolev flow.

5.5 Experiments

5.5.1 Regularity of Sobolev Active Contour

In this experiment, we show a case when the scale-invariant elastic regularity term (113) is more beneficial than the using the traditional length penalty (111). Note that the elastic regularizer does not generally have a length shrinking effect, but keeps the contour regular. This length shrinking effect may have a detrimental effect as shown in Fig. 29. Note that the length penalty restricts the curve from moving into the groves between the fingers. The elastic regularity term, on the other hand, has no such restriction, and makes the curve more smooth and rounded.

5.5.2 Tracking with Centroid/ Isoperimetric Prior

In this experiment, we illustrate one possible application of the energy (115) in tracking a man through an occlusion. For the data-based term in (115), we use the Mumford-Shah energy [72]. The prior information on the centroid and isoperimetric ratio can be obtained through a filtering process (indeed, we assume a constant

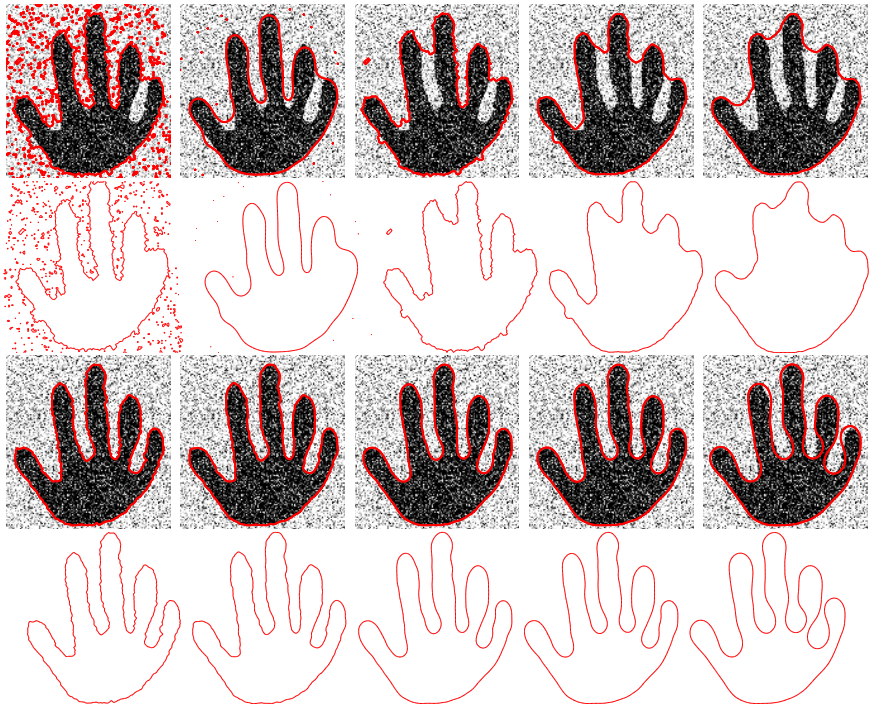


Figure 29: H^0 regularization (top two rows). Left to right: $\alpha = 1000$, $\alpha = 1000$ followed by curvature smoothing to remove the noise (least number of iterations to remove noise), $\alpha = 10000, 50000, 90000$. The image-based term is Chan-Vese. Sobolev elastic regularization (bottom two rows). Left to right: $\alpha = 0, 0.1, 5, 10, 25$. The second and fourth row show the same result as the row above them, but the image is removed for visibility.

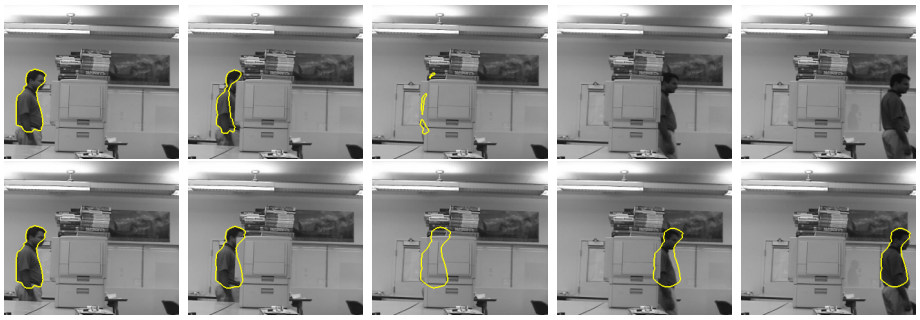


Figure 30: Tracking a man through an occlusion. Bottom row shows the results of using a prediction (filtering) on the centroid and the isoperimetric ratio, and then penalizing deviations of the contour away from predicted parameters by (115) ($\alpha = 50000$, $\beta = 100$). The top row gives the result with no such penalty. Both use Sobolev active contours.

acceleration model of both quantities). We use the tracking framework of [49] for both simulations in Fig. 30. The top row shows the result using the framework of [49] without the use of prior centroid and isoperimetric information; the bottom row incorporates this prior information. Notice that the prior information on the centroid keeps the contour moving through the occlusion, while the isoperimetric ratio (and because we are using Sobolev active contours) keeps the shape constrained.

5.5.3 Edge Detection with Non-Shrinking Model

In this experiment, we demonstrate that the traditional edge-based geodesic active contour model has an arbitrary length shrinking effect that causes the contour to pass over some meaningful edges. We show that the non-shrinking edge-based model (118) can help correct this behavior. We use edge-map, $1/(1+\phi)$, where

$$\phi(x) = \frac{1}{|B_r|}\int_{B_r(x)} (I(y) - \overline{I}_r(x))^2 \, \mathrm{d}A(y), \text{ where } \overline{I}_r(x) = \frac{1}{|B_r|}\int_{B_r(x)} I(y) \, \mathrm{d}A(y), \quad (120)$$

$B_r(x) = \{y \in \mathbb{R}^2 : \|y - x\| \leq r\}$, and $|B_r|$ denotes the area of B_r.

In Fig. 31, we segment a cyst image using various initializations. Notice that the contour with the traditional edge-based energy (using the H^0 or the Sobolev descents) consistently passes over the edge on the right side of the cyst. The non-shrinking model consistently captures the correct segmentation.

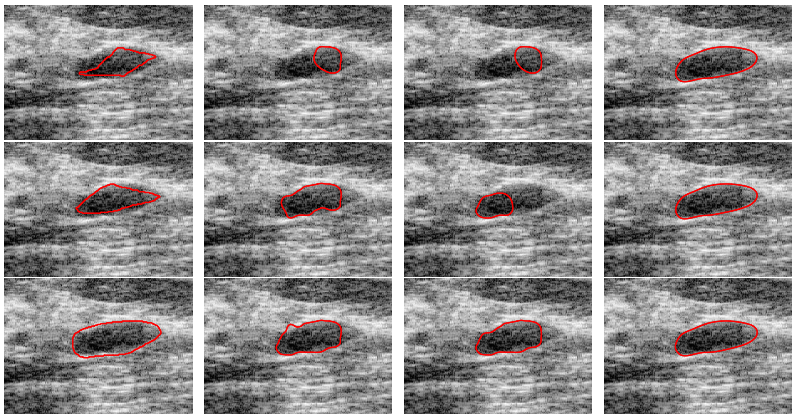


Figure 31: Segmentation of cyst image with three different initializations (first image in each row). Converged results for the (117) and H^0 active contour (second image), (117) with the Sobolev active contour (third image), and the energy (118) (last image).

5.5.4 Edge Detection by Increasing Weighted Length

In this experiment, we apply the weighted length increasing energy (119) to vessel segmentation. We show the results of using the traditional edge-based technique with a balloon; that is, we show results of using the H^0 gradient descent for the energy

$$E(c) = \int_c \phi(c(s))\,\mathrm{d}s - \alpha \int_R \phi\,\mathrm{d}A. \tag{121}$$

We use (120) as the edge-map for the weighted length increasing flow. The edge-map for (121) is $1/(1+\phi)$ where ϕ is given in (120).

In the case of vessel segmentation, it beneficial to increase the length of the initial contour more so than area. Since a vessel is characterized as a long, thin structure, a balloon term will fail to capture the global geometry of the vessel. This is demonstrated in Fig. 32: a small weight on the balloon term results in the flow capturing local features close to the initial contour; larger weights on the balloon makes the contour balloon out to capture the entire image. Note the weighted length maximizing flow does not pass the walls of the vessel since that does not increase the *length* (although it does increase area) of the contour, and is therefore able to capture the vessel.

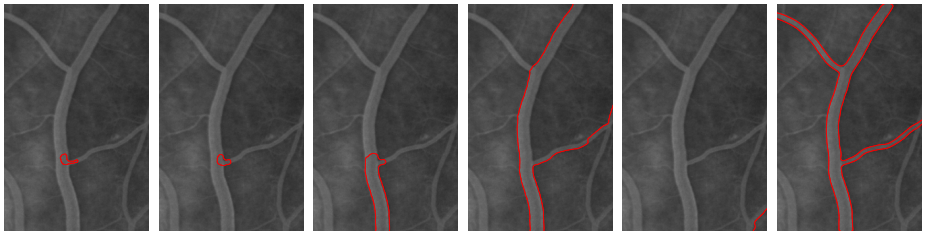


Figure 32: Left to right: initial contour, minimizing (121) $\alpha = 0.2, 0.25, 0.4$ using H^0, and increasing weighted (119) $\alpha = 0.1$ using Sobolev (all images show converged contour). The contour expands to enclose the entire image (fifth image).

5.6 Conclusion

We have demonstrated that the Sobolev gradient method allows one to consider active contour energies that were not considered in the past because the gradient method using the traditional metric cannot be used. In particular, we have given examples of energies that result in H^0 gradients that are ill-posed or high order PDEs (and hence numerically difficult to implement). These energies, as we have shown, result in Sobolev gradient flows that are both well-posed and numerically simple to implement. The experiments have shown potential uses for some energies introduced in segmentation and tracking applications.

CHAPTER VI

CONCLUSION

We have introduced novel *global* flows for optimizing active contour energies. The primary purpose of considering these novel flows was to *avoid* local minima of traditional active contour models. We have demonstrated both analytically as well as experimentally that these global flows avoid many undesirable configurations of the active contour and thereby avoid many of the local minima of traditional active contour models. As we have demonstrated, these flows have led to many applications for the object detection problem in computer vision.

We have, in Chapter 2, introduced a new class of quadratic (i.e., double) integral energies that have not been considered before in the active contour literature. These energies lead to traditional gradient flows that are global in the sense that the evolution of a point on the contour depends on all other points of the contour. We have introduced and analyzed thoroughly a specific example of this class of new flows based on electrostatic principles. This specific example was used to construct a new class of active contour evolutions that perform object detection, while also having *global regularity* and topology preserving properties. This is useful for applications where there is an a priori assumption on the geometry and topology of the object to be detected. Analytical and experimental evidence was given to support the claim that these evolutions preserve topology.

In Chapter 3, we have derived global optimizing flows for all previous active contour models by considering a new class of Sobolev-type metrics in the space of curves. We have noted that gradient flows depend not only on the energy they minimize, but also on the metric that is used in the space of curves. This fact has been ignored in

the active contour literature in the past, and the standard H^0 metric is always used to define gradient flows. We have shown a number of undesirable properties of H^0 gradient flows, and that the Sobolev-type metrics correct these problems of H^0 gradient flows. In particular, we have shown that Sobolev gradient flows are *global* flows, are less susceptible to certain local minima, are reduced order PDE when compared with H^0 gradient flows, and do not take extra computational effort when compared to H^0 gradient flows. They also produce meaningful metrics in the space of curves, and therefore our framework is a step in the direction of obtaining a *consistent* theory of shape analysis and shape optimization. In Chapter 4, we have further analyzed Sobolev gradient flows and showed that they are ideal for tracking since they have an *automatic* coarse-to-fine evolution property. Finally, in Chapter 5, we have shown that the Sobolev gradient flows not only give benefits over H^0 gradient flows, but also allow one to optimize a variety of new active contour energies that *cannot* be implemented using traditional H^0 gradient flows because of ill-posedness or numerical issues of the H^0 PDE.

This thesis has opened up a new avenue of research in the active contour field. In particular, we have shown how the metric structure on the space of shapes greatly affects the technique for optimizing energies on the space of these shapes. The Sobolev-type metrics, which we considered, illustrated this point very well. Sobolev metrics are useful for specific applications, but not all. Therefore, a direction of research is to examine other metrics on the space of shapes, and examine their applications to both shape optimization and shape analysis. Another interesting point to examine is whether the definition of shape as a region in the plane represented by its boundary contour is the best representation in order to calculate gradient flows and do calculations related to shape analysis. In particular, is there a better representation of shape that leads to meaningful and useful metrics?

In the immediate future, it is of interest to extent all of the ideas presented in

this thesis to surfaces. Specifically, we would like to extent the topology preserving flow presented in Chapter 2 to surfaces, where the problem is not only of interest for applications in computer vision, but also of great interest in the mathematics literature. Indeed, in computer vision, many researchers are interested in constructing smoothing flows that do not break the surface, unlike mean curvature flow. In the mathematics community, it is of interest to construct a flow that can retract a surface with trivial fundamental group to a sphere. The flow presented in Chapter 2 extended to surfaces may be able to achieve this retraction. In addition, we would like to extent Sobolev gradient flows to surfaces because of the many potential applications that this thesis has illustrated. One application may be tracking of heart ventricles in ultrasound 3D image sequences. Although the Sobolev norms naturally extend to surfaces, the gradients require one to solve a nonlinear Laplace equation on a surface, and this is difficult to do in a computationally feasible manner. Therefore, it may be of interest to consider modifications of these norms in order to derive simplier gradients, which have similar qualitative properties to the original Sobolev gradient flows.

APPENDIX A

CALCULATIONS FOR CHAPTER 2

A.1 Expressions for Polygon Forces

In this section, we give the expressions for $F_k^{i,j}$.

A.1.1 Non-adjacent Segments Forces

In this section, we derive the expressions for $F_k^{i,j}$. We shall assume that $C_i \cap C_j = \emptyset$, and that v_k is as shown in Fig. 33. Let ℓ_i and ℓ_j denote the lengths of segment C_i and C_j, respectively. Then, the energy contribution due to the interactions of C_i with C_j is given by $E_{i,j} = \iint_{C_i \times C_j} \frac{\mathrm{d}x\,\mathrm{d}y}{\|x-y\|}$. We shall first evaluate the inner integral. We may assume C_j lies along an x-axis with one end point at the origin, and that $(x_0, y_0) \in C_i$. Then the inner integral becomes $\int_0^{\ell_j} \frac{\mathrm{d}x}{\sqrt{(x-x_0)^2+y_0^2}}$. By choosing the change of variable $x - x_0 = y_0 \tan\phi$, one can show

$$\int_0^{\ell_j} \frac{\mathrm{d}x}{\sqrt{(x-x_0)^2+y_0^2}} = \log\left|\frac{\sqrt{(\ell_j - x_0)^2 + y_0^2} + \ell_j - x_0}{\sqrt{x_0^2+y_0^2} - x_0}\right|. \tag{122}$$

Now let $r_1 = \sqrt{(\ell_j - x_0)^2 + y_0^2}$ and $r_2 = \sqrt{x_0^2 + y_0^2}$; these are the lengths from v_j and v_{j+1} to the point (x_0, y_0). One can show with some manipulation that $-x_0 = \frac{r_2^2 - r_1^2 - \ell_j^2}{2\ell}$. Therefore, (122) becomes

$$\log\left|\frac{r_2 + \ell_j + \frac{r_2^2 - r_1^2 - \ell_j^2}{2\ell}}{r_1 + \frac{r_2^2 - r_1^2 - \ell_j^2}{2\ell}}\right| = \log\left(\frac{r_1 + r_2 + \ell_j}{r_1 + r_2 - \ell_j}\right)$$

where the absolute value becomes unnecessary because of the triangle inequality: $\ell_j \le r_1 + r_2$. Now we see that $E_{i,j}$ becomes

$$E_{i,j} = \int_0^{\ell_i} \log\left(\frac{r_1(y,\theta) + r_2(y,\theta) + \ell_j}{r_1(y,\theta) + r_2(y,\theta) - \ell_j}\right)\,\mathrm{d}y \tag{123}$$

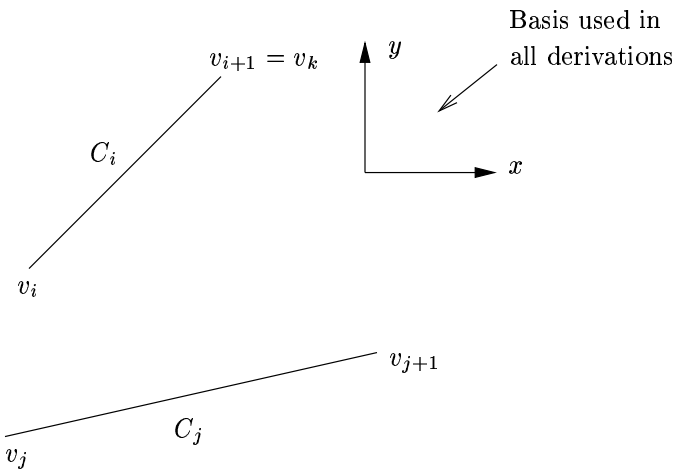


Figure 33: Configuration of segments and vertices assumed in the derivations of non-adjacent segment quantities.

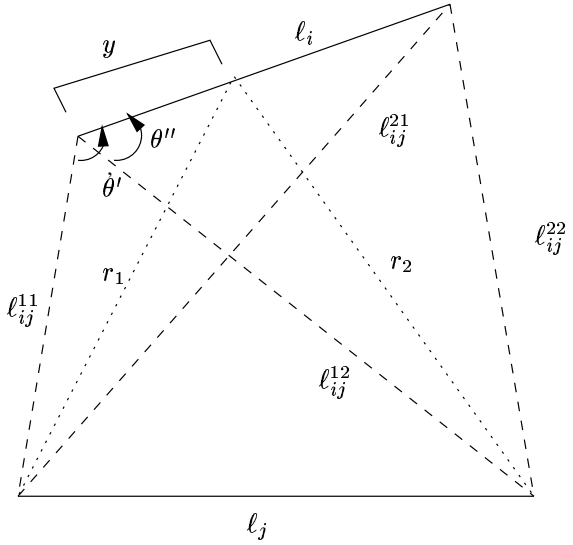


Figure 34: Names of various distances, segment lengths and angles for non-adjacent segments.

where θ is defined to be the angle between $v_{i+1} - v_i$ and $v_{j+1} - v_j$. Thus, we see

$$\frac{\partial V_{i,j}}{\partial \ell_i} = \log\left(\frac{r_1(\ell_i,\theta) + r_2(\ell_i,\theta) + \ell_j}{r_1(\ell_i,\theta) + r_2(\ell_i,\theta) - \ell_j}\right) = \log\left(\frac{\ell_{ij}^{21} + \ell_{ij}^{22} + \ell_j}{\ell_{ij}^{21} + \ell_{ij}^{22} - \ell_j}\right) \tag{124}$$

where the lengths ℓ_{ij}s are labeled in Fig. 34. The partial derivative with respect to θ is given by

$$\frac{\partial E_{i,j}}{\partial \theta} = -2\ell_j \int_0^{\ell_i} \frac{\frac{\partial}{\partial\theta}(r_1(y,\theta) + r_2(y,\theta))}{(r_1(y,\theta) + r_2(y,\theta))^2 - \ell_j^2}\,\mathrm{d}y.$$

We may now write r_1 and r_2 in terms of y and θ using the law of cosines

$$\begin{aligned} r_1^2 &= (\ell_{ij}^{11})^2 + y^2 - 2\ell_{ij}^{11}\cos\theta' \\ r_2^2 &= (\ell_{ij}^{12})^2 + y^2 - 2\ell_{ij}^{12}\cos\theta'' \end{aligned}$$

where θ' and θ'' are related to θ by a simple translation as shown in Fig. 34. Expressions for $\partial r_1/\partial\theta$ and $\partial r_2/\partial\theta$ are

$$\frac{\partial r_1}{\partial\theta} = \frac{\ell_{ij}^{11} y\sin\theta'}{r_1} \quad \text{and,} \quad \frac{\partial r_2}{\partial\theta} = \frac{\ell_{ij}^{12} y\sin\theta''}{r_2}.$$

Therefore, we arrive at

$$\frac{\partial V_{i,j}}{\partial\theta} = -2\ell_j\left[\ell_{ij}^{11}\sin\theta'\int_0^{\ell_i}\frac{y}{r_1\left[(r_1+r_2)^2 - \ell_j^2\right]}\,\mathrm{d}y + \ell_{ij}^{12}\sin\theta''\int_0^{\ell_i}\frac{y}{r_2\left[(r_1+r_2)^2-\ell_j^2\right]}\,\mathrm{d}y\right] \tag{125}$$

after substituting expressions derived above. This integral cannot be evaluated explicitly; therefore, a numerical integration must be performed in the implementation.

We now relate $\partial E_{i,j}/\partial\ell_i$ and $\partial E_{i,j}/\partial\theta$ to $\partial E_{i,j}/\partial v_k$ where we have assumed v_k is as shown in Fig. 33. Note that there is no loss of generality by assuming this. First, we note that

$$v_k = \ell_i\begin{pmatrix}\cos\theta\\ \sin\theta\end{pmatrix} + v_i$$

therefore, by the chain rule, we have

$$\begin{pmatrix}\frac{\partial E_{i,j}}{\partial\ell_i} & \frac{\partial E_{i,j}}{\partial\theta}\end{pmatrix}(\ell_i,\theta) = \mathbf{D}E_{i,j}(v_k(\ell_i,\theta))\circ\mathbf{D}v_k(\ell_i,\theta).$$

Thus, we can readily derive that

$$F_k^{i,j} = -\frac{\partial E_{i,j}}{\partial v_k} = -\frac{1}{\ell_i}\begin{pmatrix} \ell_i\cos\theta & \sin\theta \\ \ell_i\sin\theta & -\cos\theta \end{pmatrix}\begin{pmatrix} \frac{\partial E_{i,j}}{\partial \ell_i} \\ \frac{\partial E_{i,j}}{\partial \theta} \end{pmatrix} \tag{126}$$

where $\partial E_{i,j}/\partial \ell_i$ and $\partial E_{i,j}/\partial \theta$ are given in (124) and (125), respectively.

A.1.2 Adjacent Segments Forces

In this section, we derive the expressions for $F_i^{i-1,i}$, $F_{i-1}^{i-1,i}$, and $F_{i+1}^{i-1,i}$. Let θ be the adjacent angle between C_j and C_{j-1}. Then $E_{i-1,i}$, the energy contribution due to the interactions of C_{i-1} on C_i, is given by

$$E_{i-1,i} = \iint_{C_{i-1}\times C_i} \frac{\mathrm{d}x\,\mathrm{d}y}{\|x-y\|} = \int_0^{\ell_{i-1}}\int_0^{\ell_i} \frac{\mathrm{d}x\,\mathrm{d}y}{\sqrt{x^2+y^2-2xy\cos\theta}}$$

using an application of the law of cosines. Expanding the inner integral, we have

$$\begin{aligned} E_{i-1,i} &= \int_0^{\ell_{i-1}} \log\left(x - y\cos\theta + \sqrt{x^2+y^2-2xy\cos\theta}\right)\Big|_0^{\ell_i}\,\mathrm{d}y \\ &= \Big[\Big[-y + x\log\left(y - x\cos\theta + \sqrt{x^2+y^2-2xy\cos\theta}\right) \\ &+ \quad y\log\left(x - y\cos\theta + \sqrt{x^2+y^2-2xy\cos\theta}\right)\Big]_0^{\ell_i}\Big]_0^{\ell_{i-1}} \\ &= \ell_i\log\left(\ell_{i-1} - \ell_i\cos\theta + \ell_{i-1i}^{22}\right) + \ell_{i-1}\log\left(\ell_i - \ell_{i-1}\cos\theta + \ell_{i-1i}^{22}\right) \\ &- \ell_i\log\left(-\ell_i\cos\theta + \ell_i\right) - \ell_{i-1}\log\left(-\ell_{i-1}\cos\theta + \ell_{i-1}\right) \end{aligned}$$

where we have noted that $(\ell_{i-1i}^{22})^2 = \ell_{i-1}^2 + \ell_i^2 - 2\ell_{i-1}\ell_i\cos\theta$. Now we substitute $-\cos\theta = ((\ell_{i-1i}^{22})^2 - \ell_{i-1}^2 - \ell_i^2)/(2\ell_{i-1}\ell_i)$ into the above expression to find

$$\begin{aligned} E_{i-1,i} &= \ell_i\log\left(\frac{\ell_{i-1} + \ell_{i-1i}^{22} + \frac{(\ell_{i-1i}^{22})^2-\ell_{i-1}^2-\ell_i^2}{2\ell_{i-1}}}{\ell_i + \frac{(\ell_{i-1i}^{22})^2-\ell_{i-1}^2-\ell_i^2}{2\ell_{i-1}}}\right) \\ &+ \ell_{i-1}\log\left(\frac{\ell_i + \ell_{i-1i}^{22} + \frac{(\ell_{i-1i}^{22})^2-\ell_{i-1}^2-\ell_i^2}{2\ell_i}}{\ell_{i-1} + \frac{(\ell_{i-1i}^{22})^2-\ell_{i-1}^2-\ell_i^2}{2\ell_i}}\right) \\ &= \ell_i\log\left(\frac{(\ell_{i-1}+\ell_{i-1i}^{22})^2 - \ell_i^2}{(\ell_{i-1i}^{22})^2 - (\ell_{i-1}-\ell_i)^2}\right) + \ell_{i-1}\log\left(\frac{(\ell_i+\ell_{i-1i}^{22})^2 - \ell_{i-1}^2}{(\ell_{i-1i}^{22})^2 - (\ell_{i-1}-\ell_i)^2}\right) \\ &= \ell_i\log\left(\frac{\ell_{i-1i}^{22}+\ell_i+\ell_{i-1}}{\ell_{i-1i}^{22}+\ell_i-\ell_{i-1}}\right) + \ell_{i-1}\log\left(\frac{\ell_{i-1i}^{22}+\ell_{i-1}+\ell_i}{\ell_{i-1i}^{22}+\ell_{i-1}-\ell_i}\right). \end{aligned} \tag{127}$$

We now note that

$$\ell_{i-1} = \|v_{i-1} - v_i\|, \ell_i = \|v_i - v_{i+1}\|, \text{ and } \ell^{22}_{i-1i} = \|v_{i-1} - v_{i+1}\| \tag{128}$$

and therefore,

$$\frac{\partial \ell_{i-1}}{\partial v_i} = \frac{v_i - v_{i-1}}{\ell_{i-1}} \tag{129}$$

and the other partial derivatives of the lengths defined in (128) are similar to above. We can now compute $\partial E_{i-1,i}/\partial v_i$; after manipulation, we arrive at

$$\begin{aligned}
\frac{\partial E_{i-1,i}}{\partial v_i} &= \frac{\partial \ell_i}{\partial v_i} \log\left(\frac{\ell^{22}_{i-1i} + \ell_i + \ell_{i-1}}{\ell^{22}_{i-1i} + \ell_i - \ell_{i-1}}\right) + 2\ell_i \frac{(\ell^{22}_{i-1i} + \ell_i)\frac{\partial \ell_{i-1}}{\partial v_i} - \ell_{i-1}\frac{\partial \ell_i}{\partial v_i}}{(\ell^{22}_{i-1i} + \ell_i)^2 - \ell^2_{i-1}} \\
&+ \frac{\partial \ell_{i-1}}{\partial v_i} \log\left(\frac{\ell^{22}_{i-1i} + \ell_{i-1} + \ell_i}{\ell^{22}_{i-1i} + \ell_{i-1} - \ell_i}\right) + 2\ell_{i-1} \frac{(\ell^{22}_{i-1i} + \ell_{i-1})\frac{\partial \ell_i}{\partial v_i} - \ell_i \frac{\partial \ell_{i-1}}{\partial v_i}}{(\ell^{22}_{i-1i} + \ell_{i-1})^2 - \ell^2_i}
\end{aligned}$$

Hence, by substituting expressions for the partial derivatives of the lengths that are given in (129), we have

$$\begin{aligned}
F_i^{i-1,i} &= -\frac{v_i - v_{i+1}}{\ell_i} \log\left(\frac{\ell^{22}_{i-1i} + \ell_i + \ell_{i-1}}{\ell^{22}_{i-1i} + \ell_i - \ell_{i-1}}\right) \\
&- 2\ell_i \frac{\frac{\ell^{22}_{i-1i} + \ell_i}{\ell_{i-1}}(v_i - v_{i-1}) - \frac{\ell_{i-1}}{\ell_i}(v_i - v_{i+1})}{(\ell^{22}_{i-1i} + \ell_i)^2 - \ell^2_{i-1}} \\
&- \frac{v_i - v_{i-1}}{\ell_{i-1}} \log\left(\frac{\ell^{22}_{i-1i} + \ell_{i-1} + \ell_i}{\ell^{22}_{i-1i} + \ell_{i-1} - \ell_i}\right) \\
&- 2\ell_{i-1} \frac{\frac{\ell^{22}_{i-1i} + \ell_{i-1}}{\ell_i}(v_i - v_{i+1}) - \frac{\ell_i}{\ell_{i-1}}(v_i - v_{i-1})}{(\ell^{22}_{i-1i} + \ell_{i-1})^2 - \ell^2_i}.
\end{aligned} \tag{130}$$

Similarly, we can compute $F_{i-1}^{i-1,i} = -\partial E_{i-1,i}/\partial v_{i-1}$. The resulting expression is

$$\begin{aligned}
F_{i-1}^{i-1,i} &= 2\ell_i \frac{\frac{\ell_{i-1}}{\ell^{22}_{i-1i}}(v_{i-1} - v_{i+1}) - \frac{\ell^{22}_{i-1i} + \ell_i}{\ell_{i-1}}(v_{i-1} - v_i)}{(\ell^{22}_{i-1i} + \ell_i)^2 - \ell^2_{i-1}} \\
&- \frac{v_{i-1} - v_i}{\ell_{i-1}} \log\left(\frac{\ell^{22}_{i-1i} + \ell_{i-1} + \ell_i}{\ell^{22}_{i-1i} + \ell_{i-1} - \ell_i}\right) \\
&+ 2\ell_{i-1}\ell_i \frac{\frac{v_{i-1} - v_{i+1}}{\ell^{22}_{i-1i}} + \frac{v_{i-1} - v_i}{\ell_{i-1}}}{(\ell^{22}_{i-1i} + \ell_{i-1})^2 - \ell^2_i}
\end{aligned} \tag{131}$$

and $F_{i+1}^{i-1,i} = -\partial E_{i-1,i}/\partial v_{i+1}$ is obtained simply by switching v_{i-1} with v_{i+1}, and vice-versa in (131).

A.2 Computation of ∇E_c

A.2.1 Notation

Define $f : \mathbb{R} \to \mathbb{R}$ as $f(x) = x^{-1}$, and $\phi : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ as $\phi(x, y) = f(\|x - y\|)$. Let $p \in S^1$ then define $B_p(\epsilon) = \{q \in S^1 : d_{S^1}(p, q) < \epsilon\}$ where $d_{S^1} : S^1 \times S^1 \to [0, 1/2]$ is the distance on S^1 defined by

$$d_{S^1}(p, q) = \begin{cases} \min\{p - q, 1 + q - p\} & \text{for } p > q \\ \min\{q - p, 1 + p - q\} & \text{for } p \leq q \end{cases}$$

If $C \in C^2(S^1, \mathbb{R}^2)$ is a parametrization of a curve, then the geodesic distance is $d_C : S^1 \times S^1 \to \mathbb{R}^+$, which is defined by

$$d_C(p, q) = \begin{cases} \min\{\int_q^p \|C'(\xi)\| \, \mathrm{d}\xi, \int_p^1 \|C'(\xi)\| \, \mathrm{d}\xi + \int_0^q \|C'(\xi)\| \, \mathrm{d}\xi\} & \text{for } p > q \\ \min\{\int_p^q \|C'(\xi)\| \, \mathrm{d}\xi, \int_q^1 \|C'(\xi)\| \, \mathrm{d}\xi + \int_0^p \|C'(\xi)\| \, \mathrm{d}\xi\} & \text{for } p \leq q \end{cases}. \quad (132)$$

Define $E^\epsilon : C^2(S^1, \mathbb{R}^2) \to \mathbb{R}^+$ as

$$E^\epsilon(C) = E_1^\epsilon(C) - E_2^\epsilon(C) \quad (133)$$

where $E_1^\epsilon, E_2^\epsilon : C^2(S^1, \mathbb{R}^2) \to \mathbb{R}^+$ are defined as

$$E_1^\epsilon(C) = \frac{1}{2} \int_{S^1} \int_{S^1 \setminus B_p(\epsilon)} \phi(C(p), C(q)) \|C'(q)\| \|C'(p)\| \, \mathrm{d}q \, \mathrm{d}p \quad (134)$$

$$E_2^\epsilon(C) = \frac{1}{2} \int_{S^1} \int_{S^1 \setminus B_p(\epsilon)} f(d_C(p, q)) \|C'(q)\| \|C'(p)\| \, \mathrm{d}q \, \mathrm{d}p. \quad (135)$$

We note that $\mathrm{d}p$, $\mathrm{d}q$, $\mathrm{d}\xi$ are push forwards of Lebesgue measure onto S^1. $\mathbf{D}_1\phi$ and $\mathbf{D}_2\phi$ will denote the first and second partial derivatives of ϕ.

A.2.2 Overview of Computation

First, it is easy to see that $E_c(C) = \lim_{\epsilon \to 0} E^\epsilon(C)$. It is important to notice that $E^\epsilon(C)$ depends on the parametrization chosen for a curve, but the limit does not. The computation goes as follows. First, $\mathrm{d}E^\epsilon(C) \cdot \tilde{C}$ is computed where $C \in C^2(S^1, \mathbb{R}^2)$. Then, some useful estimates of quantities of the form $\phi - f$ will be derived. These estimates will help compute $\lim_{\epsilon \to 0} \mathrm{d}E^\epsilon(C) \cdot \tilde{C}$. Finally, we show that $\mathrm{d}E_c(C) \cdot \tilde{C} = \lim_{\epsilon \to 0} \mathrm{d}E^\epsilon(C) \cdot \tilde{C}$, which will finish the computation.

A.2.3 Computation of $\mathrm{d}E^\epsilon$

We first compute the first variation of E_1^ϵ:

$$2\,\mathrm{d}E_1^\epsilon(C)\cdot\tilde{C} = \int_{S^1}\int_{S^1\setminus B_p(\epsilon)} \mathbf{D}_1\phi(C(p),C(q))\cdot\tilde{C}(p)\|C'(q)\|\|C'(p)\|\,\mathrm{d}q\,\mathrm{d}p \tag{136}$$

$$+\int_{S^1}\int_{S^1\setminus B_p(\epsilon)} \mathbf{D}_2\phi(C(p),C(q))\cdot\tilde{C}(q)\|C'(q)\|\|C'(p)\|\,\mathrm{d}q\,\mathrm{d}p \tag{137}$$

$$+\int_{S^1}\int_{S^1\setminus B_p(\epsilon)} \phi(C(p),C(q))\frac{\tilde{C}'(q)\cdot C'(q)}{\|C'(q)\|}\|C'(p)\|\,\mathrm{d}q\,\mathrm{d}p \tag{138}$$

$$+\int_{S^1}\int_{S^1\setminus B_p(\epsilon)} \phi(C(p),C(q))\frac{\tilde{C}'(p)\cdot C'(p)}{\|C'(p)\|}\|C'(q)\|\,\mathrm{d}q\,\mathrm{d}p. \tag{139}$$

We note that $\mathbf{D}_2\phi(x,y)=\mathbf{D}_1\phi(y,x)$ by symmetry of ϕ. Note that in (137) and (138) that the integrands are bounded, and the measures are finite. Thus, by Fubini's Theorem, we may switch order of integration. Switching order of integration and substituting p for q and q for p in (137) and (138), we arrive at

$$\mathrm{d}E_1^\epsilon(C)\cdot\tilde{C} = \int_{S^1}\left\langle \tilde{C}(p), \int_{S^1\setminus B_p(\epsilon)} \mathbf{D}_1\phi(C(p),C(q))\|C'(q)\|\,\mathrm{d}q\right\rangle\|C'(p)\|\,\mathrm{d}p \tag{140}$$

$$+\int_{S^1}\left\langle \tilde{C}'(p), \int_{S^1\setminus B_p(\epsilon)} \phi(C(p),C(q))\|C'(q)\|\,\mathrm{d}q\,\mathcal{T}(p)\right\rangle\,\mathrm{d}p. \tag{141}$$

Integrating (141) by parts we see that

$$\mathrm{d}E_1^\epsilon(C)\cdot\tilde{C} = \int_{S^1}\left\langle \tilde{C}(p), \int_{S^1\setminus B_p(\epsilon)} \mathbf{D}_1\phi(C(p),C(q))\|C'(q)\|\,\mathrm{d}q\right\rangle\|C'(p)\|\,\mathrm{d}p \tag{142}$$

$$-\int_{S^1}\left\langle \tilde{C}(p), \int_{S^1\setminus B_p(\epsilon)} \phi(C(p),C(q))\|C'(q)\|\,\mathrm{d}q\kappa(p)\mathcal{N}(p)\right\rangle\|C'(p)\|\,\mathrm{d}p \tag{143}$$

$$-\int_{S^1}\left\langle \tilde{C}(p), \int_{S^1\setminus B_p(\epsilon)} \mathbf{D}_1\phi(C(p),C(q))\cdot C'(p)\|C'(q)\|\,\mathrm{d}q\,\mathcal{T}(p)\right\rangle\,\mathrm{d}p \tag{144}$$

$$-\int_{S^1}\Big\langle \tilde{C}(p), [\phi(C(p),C(p-\epsilon))\|C'(p-\epsilon)\| - \phi(C(p),C(p+\epsilon))\|C'(p+\epsilon)\|]\,\mathcal{T}(p)\Big\rangle\,\mathrm{d}p. \tag{145}$$

Combining (142) and (144) we find

$$\mathrm{d}E_1^\epsilon(C)\cdot\tilde{C} = \int_{S^1}\left\langle \tilde{C}(p), \int_{S^1\setminus B_p(\epsilon)} \mathbf{D}_1\phi(C(p),C(q))\cdot\mathcal{N}(p)\|C'(q)\|\,\mathrm{d}q\,\mathcal{N}(p)\right\rangle \cdot\|C'(p)\|\,\mathrm{d}p \tag{146}$$

$$-\int_{S^1}\left\langle \tilde{C}(p), \int_{S^1\setminus B_p(\epsilon)} \phi(C(p),C(q))\|C'(q)\|\,\mathrm{d}q\kappa(p)\mathcal{N}(p)\right\rangle \|C'(p)\|\,\mathrm{d}p \tag{147}$$

$$-\int_{S^1}\Big\langle \tilde{C}(p), [\phi(C(p),C(p-\epsilon))\|C'(p-\epsilon)\| -\phi(C(p),C(p+\epsilon))\|C'(p+\epsilon)\|]\,\mathcal{T}(p)\Big\rangle\,\mathrm{d}p. \tag{148}$$

Now we calculate the first variation of E_2^ϵ. Let us note that we may write E_2^ϵ as

$$E_2^\epsilon(C) = \int_{S^1} F(p,C)\|C'(p)\|\,\mathrm{d}p \tag{149}$$

where $F : S^1\times C^2(S^1,\mathbb{R}^2)\to\mathbb{R}^+$ is given by

$$F(p,C) = \int_{S^1\setminus B_p(\epsilon)} f(d_C(p,q))\|C'(q)\|\,\mathrm{d}q. \tag{150}$$

Computing the first variation of E_2^ϵ in terms of F we see that

$$\mathrm{d}E_2^\epsilon(C)\cdot\tilde{C} = \int_{S^1}\delta F(p,C)\cdot\tilde{C}\|C'(p)\|\,\mathrm{d}p + \int_{S^1} F(p,C)\frac{C'(p)\cdot\tilde{C}'(p)}{\|C'(p)\|}\,\mathrm{d}p. \tag{151}$$

Integrating the second term in (151) by parts we see that

$$\mathrm{d}E_2^\epsilon(C)\cdot\tilde{C} = \int_{S^1}\delta F(p,C)\cdot\tilde{C}\|C'(p)\|\,\mathrm{d}p \tag{152}$$

$$-\int_{S^1}\left\langle \tilde{C}(p), \frac{\mathrm{d}}{\mathrm{d}p}\left[F(p,C)\frac{C'(p)}{\|C'(p)\|}\right]\right\rangle\,\mathrm{d}p \tag{153}$$

$$= \int_{S^1}\delta F(p,C)\cdot\tilde{C}\|C'(p)\|\,\mathrm{d}p - \int_{S^1}\Big\langle \tilde{C}(p), F(p,C)\kappa(p)\mathcal{N}(p)\Big\rangle\|C'(p)\|\,\mathrm{d}p - \int_{S^1}\Big\langle \tilde{C}(p), F_p(p,C)\mathcal{T}(p)\Big\rangle\,\mathrm{d}p. \tag{154}$$

We now compute $\delta F(p,C)\cdot\tilde{C}$ and $F_p(p,C)$. But we first note that by the change of variable $u = d_C(p,q)$ where p is fixed, F can be written in the form

$$F(p,C) = \int_{d_C(p,p-\epsilon)}^{\frac{L}{2}} f(u)\,\mathrm{d}u + \int_{d_C(p,p+\epsilon)}^{\frac{L}{2}} f(u)\,\mathrm{d}u \tag{155}$$

where L denotes the length of C. Therefore,

$$\delta F(p,C)\cdot\tilde{C} = f(L/2)\delta L(C)\cdot\tilde{C} - f(d_C(p,p-\epsilon))\delta d_C(p,p-\epsilon)\cdot\tilde{C} - f(d_C(p,p+\epsilon))\delta d_C(p,p+\epsilon)\cdot\tilde{C} \tag{156}$$

and

$$F_p(p,C) = -f(d_C(p,p-\epsilon))(\|C'(p)\| - \|C'(p-\epsilon)\|) - f(d_C(p,p+\epsilon))(\|C'(p+\epsilon)\| - \|C'(p)\|) \tag{157}$$

where we note that

$$\delta L(C)\cdot\tilde{C} = -\int_{S^1}\left\langle\tilde{C}(q),\kappa(q)\mathcal{N}(q)\right\rangle\|C'(q)\|\,\mathrm{d}q \tag{158}$$

$$\delta d_C(p,p-\epsilon)\cdot\tilde{C} = \left\langle\tilde{C}(q),\mathcal{T}(q)\right\rangle\Big|_{p-\epsilon}^{p} - \int_{p-\epsilon}^{p}\left\langle\tilde{C}(q),\kappa(q)\mathcal{N}(q)\right\rangle\|C'(q)\|\,\mathrm{d}q \tag{159}$$

$$\delta d_C(p,p+\epsilon)\cdot\tilde{C} = \left\langle\tilde{C}(q),\mathcal{T}(q)\right\rangle\Big|_{p}^{p+\epsilon} - \int_{p}^{p+\epsilon}\left\langle\tilde{C}(q),\kappa(q)\mathcal{N}(q)\right\rangle\|C'(q)\|\,\mathrm{d}q. \tag{160}$$

We now simplify the expression in (152) by substituting the expressions in (156) and (157). Let us first consider the tangent terms:

$$\begin{aligned}-\int_{S^1} f(d_C(p,p-\epsilon))\left\langle\tilde{C}(q),\mathcal{T}(q)\right\rangle\Big|_{p-\epsilon}^{p}\|C'(p)\|\,\mathrm{d}p \\ -\int_{S^1} f(d_C(p,p+\epsilon))\left\langle\tilde{C}(q),\mathcal{T}(q)\right\rangle\Big|_{p}^{p+\epsilon}\|C'(p)\|\,\mathrm{d}p + \\ \int_{S^1}\left\langle\tilde{C}(p), f(d_C(p,p-\epsilon))(\|C'(p)\| - \|C'(p-\epsilon)\|)\mathcal{T}(p)\right\rangle\,\mathrm{d}p + \\ \int_{S^1}\left\langle\tilde{C}(p), f(d_C(p,p+\epsilon))(\|C'(p+\epsilon)\| - \|C'(p)\|)\mathcal{T}(p)\right\rangle\,\mathrm{d}p.\end{aligned} \tag{161}$$

This simplifies to

$$2\int_{S^1}\left\langle\tilde{C}(p),\left[f(d_C(p,p+\epsilon))\|C'(p+\epsilon)\| - f(d_C(p,p-\epsilon))\|C'(p-\epsilon)\|\right]\mathcal{T}(p)\right\rangle\,\mathrm{d}p. \tag{162}$$

Putting this together with the normal terms and combining this with (146), we arrive at

$$\mathrm{d}E^{\epsilon}(C)\cdot\tilde{C} = 2\int_{S^1}\left\langle \tilde{C}(p), \int_{S^1\setminus B_p(\epsilon)} \mathbf{D}_1\phi(C(p),C(q))\cdot\mathcal{N}(p)\|C'(q)\|\,\mathrm{d}q\mathcal{N}(p)\right\rangle \cdot\|C'(p)\|\,\mathrm{d}p \tag{163}$$

$$-2\int_{S^1}\left\langle \tilde{C}(p), \int_{S^1\setminus B_p(\epsilon)} \phi(C(p),C(q))\|C'(q)\|\,\mathrm{d}q\kappa(p)\mathcal{N}(p)\right\rangle\|C'(p)\|\,\mathrm{d}p \tag{164}$$

$$+\int_{S^1}\left\langle \tilde{C}(p), Lf(L/2)\kappa(p)\mathcal{N}(p)\right\rangle\|C'(p)\|\,\mathrm{d}p \tag{165}$$

$$-\int_{S^1} f(d_C(p,p-\epsilon))\int_{p-\epsilon}^{p}\left\langle \tilde{C}(q),\kappa(q)\mathcal{N}(q)\right\rangle\|C'(q)\|\,\mathrm{d}q\|C'(p)\|\,\mathrm{d}p \tag{166}$$

$$-\int_{S^1} f(d_C(p,p+\epsilon))\int_{p}^{p+\epsilon}\left\langle \tilde{C}(q),\kappa(q)\mathcal{N}(q)\right\rangle\|C'(q)\|\,\mathrm{d}q\|C'(p)\|\,\mathrm{d}p \tag{167}$$

$$+\int_{S^1}\left\langle \tilde{C}(p), \int_{S^1\setminus B_p(\epsilon)} f(d_C(p,q))\|C'(q)\|\,\mathrm{d}q\kappa(p)\mathcal{N}(p)\right\rangle\|C'(p)\|\,\mathrm{d}p \tag{168}$$

$$-2\int_{S^1}\Big\langle \tilde{C}(p), [\phi(C(p),C(p-\epsilon)) - f(d_C(p,p-\epsilon))]\,\|C'(p-\epsilon)\|\mathcal{T}(p)\Big\rangle\,\mathrm{d}p \tag{169}$$

$$+2\int_{S^1}\Big\langle \tilde{C}(p), [\phi(C(p),C(p+\epsilon)) -f(d_C(p,p+\epsilon))]\,\|C'(p+\epsilon)\|\mathcal{T}(p)\Big\rangle\,\mathrm{d}p. \tag{170}$$

A.2.4 Limit Variation

We now compute $\lim_{\epsilon\downarrow 0} \mathrm{d}E^{\epsilon}(C)\cdot\tilde{C}$. Let us note that by Lebesgue's Differentiation Theorem,

$$\lim_{\epsilon\to 0}\frac{\int_{p-\epsilon}^{p}\left\langle\tilde{C}(q),\kappa(q)\mathcal{N}(q)\right\rangle\|C'(q)\|\,\mathrm{d}q}{d_C(p,p-\epsilon)}=\langle C(p),\kappa(p)\mathcal{N}(p)\rangle\,,$$
$$\lim_{\epsilon\to 0}\frac{\int_{p}^{p+\epsilon}\left\langle\tilde{C}(q),\kappa(q)\mathcal{N}(q)\right\rangle\|C'(q)\|\,\mathrm{d}q}{d_C(p,p+\epsilon)}=\langle C(p),\kappa(p)\mathcal{N}(p)\rangle\,.$$

Since $\tilde{C}\in L^2$, and C' and κ are bounded (they are continuous functions on a compact set), and application of Hölder's inequality, we have

$$\sup_{p\in S^1}\left|\frac{\int_{p-\epsilon}^{p}\left\langle\tilde{C}(q),\kappa(q)\mathcal{N}(q)\right\rangle\|C'(q)\|\,\mathrm{d}q}{d_C(p,p-\epsilon)}\right|<M$$
$$\sup_{p\in S^1}\left|\frac{\int_{p}^{p+\epsilon}\left\langle\tilde{C}(q),\kappa(q)\mathcal{N}(q)\right\rangle\|C'(q)\|\,\mathrm{d}q}{d_C(p,p+\epsilon)}\right|<M$$

and so by the Lebesgue Dominated Convergence Theorem,

$$\begin{aligned}
&\int_{S^1}\left\langle\tilde{C}(p),\kappa(p)\mathcal{N}(p)\right\rangle\|C'(p)\|\,\mathrm{d}p\\
&\quad=\lim_{\epsilon\to 0}\int_{S^1}f(d_C(p,p-\epsilon))\int_{p-\epsilon}^{p}\left\langle\tilde{C}(q),\kappa(q)\mathcal{N}(q)\right\rangle\|C'(q)\|\,\mathrm{d}q\|C'(p)\|\,\mathrm{d}p,\\
&\int_{S^1}\left\langle\tilde{C}(p),\kappa(p)\mathcal{N}(p)\right\rangle\|C'(p)\|\,\mathrm{d}p\\
&\quad=\lim_{\epsilon\to 0}\int_{S^1}f(d_C(p,p+\epsilon))\int_{p}^{p+\epsilon}\left\langle\tilde{C}(q),\kappa(q)\mathcal{N}(q)\right\rangle\|C'(q)\|\,\mathrm{d}q\|C'(p)\|\,\mathrm{d}p.
\end{aligned}$$

Therefore, in the limit, (165)-(167) combine to give 0 since $Lf(L/2)=2$.

We now compute the limit of the terms involving tangent components (169) and (170). By Lemma A.3.1 in Appendix A.3, we have that $\phi(C(s+t),C(s))=1/t+ct$ where we have abused notation using an argument of arclength in C. Switching back to the parametric domain, and substituting $t=d_C(p,p+\epsilon)$ and $t=d_C(p,p-\epsilon)$, respectively, we have that

$$\begin{aligned}
\phi(C(p),C(p+\epsilon))-f(d_C(p,p+\epsilon)) &= c_1(p)d_C(p,p+\epsilon)\\
\phi(C(p),C(p-\epsilon))-f(d_C(p,p-\epsilon)) &= c_2(p)d_C(p,p-\epsilon).
\end{aligned}$$

where $c_1, c_2 \in C(S^1, \mathbb{R}^+)$. Thus, clearly we see that

$$\lim_{\epsilon\to 0} \phi(C(p), C(p+\epsilon)) - f(d_C(p, p+\epsilon)) = \lim_{\epsilon\to 0} \phi(C(p), C(p-\epsilon)) - f(d_C(p, p-\epsilon)) = 0.$$

Also, since $c_1 d_C(\cdot, \cdot + \epsilon)$ and $c_2 d_C(\cdot, \cdot - \epsilon)$ are continuous functions on a compact set, they are bounded. Thus, by the Lebesgue Dominated Convergence Theorem,

$$\lim_{\epsilon\to 0} \int_{S^1} \left\langle \tilde{C}(p), [\phi(C(p), C(p-\epsilon)) - f(d_C(p, p-\epsilon))] \, \|C'(p-\epsilon)\| \mathcal{T}(p) \right\rangle \, \mathrm{d}p = 0$$
$$\lim_{\epsilon\to 0} \int_{S^1} \left\langle \tilde{C}(p), [\phi(C(p), C(p+\epsilon)) - f(d_C(p, p+\epsilon))] \, \|C'(p+\epsilon)\| \mathcal{T}(p) \right\rangle \, \mathrm{d}p = 0.$$

By combining all the previous results, we see that

$$\lim_{\epsilon\to 0} \mathrm{d}E^\epsilon(C) \cdot \tilde{C} = \lim_{\epsilon\to 0}$$
$$\left[-2 \int_{S^1} \left\langle \tilde{C}(p), \int_{S^1 \setminus B_p(\epsilon)} \frac{C(p) - C(q)}{\|C(p) - C(q)\|^3} \cdot \mathcal{N}(p) \|C'(q)\| \, \mathrm{d}q \, \mathcal{N}(p) \right\rangle \|C'(p)\| \, \mathrm{d}p \right.$$
$$- 2 \int_{S^1} \left\langle \tilde{C}(p), \int_{S^1 \setminus B_p(\epsilon)} \frac{1}{\|C(p) - C(q)\|} \|C'(q)\| \, \mathrm{d}q \, \kappa(p) \mathcal{N}(p) \right\rangle \|C'(p)\| \, \mathrm{d}p$$
$$\left. + \int_{S^1} \left\langle \tilde{C}(p), \int_{S^1 \setminus B_p(\epsilon)} \frac{1}{d_C(p, q)} \|C'(q)\| \, \mathrm{d}q \kappa(p) \mathcal{N}(p) \right\rangle \|C'(p)\| \, \mathrm{d}p \right]$$

A.3 Estimates

Lemma A.3.1. *If $C : S^1 \to \mathbb{R}^2$ is smooth and C is parametrized by arclength then*

$$\frac{\kappa(s)}{\|C(s) - C(s+t)\|} = \kappa(s) t^{-1} - \frac{1}{8} \kappa^3(s) t + \mathcal{O}(t^2)$$

and

$$\frac{(C(s) - C(s+t)) \cdot \mathcal{N}(s)}{\|C(s) - C(s+t)\|^3} = -\frac{1}{2} \kappa(s) t^{-1} - \frac{1}{6} \kappa'(s) + \mathcal{O}(t)$$

for t small. Note κ is the curvature and $\mathcal{N}$ is the inward unit normal to C.

Proof. The proof of the first estimate is essentially given in [77]. For the second estimate, using that $f(x) = (1+x)^{-3/2} = 1 - 3/2x + o(x)$ and that $\|C(s+t) - C(s)\|^2 = t^2(1 - 1/12\kappa^2(s)t^2 + o(t^2))$ we find $\|C(s+t) - C(s)\|^{-3} = t^{-3}(1 + 1/8\kappa^2(s)t^2 + o(t^2))$. Also, it is easy to see that $(C(s) - C(s+t)) \cdot \mathcal{N}(s) = -t^2/2(\kappa(s) + 1/3\kappa'(s)t + o(t))$. Multiplying the last two estimates, gives us the second result of the lemma. □

REFERENCES

[1] Abrams, A., Cantarella, J., Fu, J., Ghomi, M., and Howard, R., "Circles minimize most knot energies," *Topology*, no. 42, pp. 381–394, 2003.

[2] Adalsteinsson, D. and Sethian, J., "A fast level set method for propagating interfaces," *J. of Comp. Phys.*, vol. 118, pp. 269–277, 1995.

[3] Alexandrov, O. and Santosa, F., "A topology-preserving level set method for shape optimization," *J. Comp. Phys.*, vol. 204, pp. 121–130, 2005.

[4] Bertrand, G., "Simple points, topological numbers and geodesic neighborhoods in cubic grids," *Pattern Recognition Letters*, vol. 15, pp. 1003–1011, 1994.

[5] Blake, A. and Brockett, R., "On snakes and estimation theory," in *IEEE CDC*, 1994.

[6] Blake, A. and Isard, M., *Active Contours*. Springer Verlag, 1998.

[7] Blake, A. and Zisserman, A., "Using weak continuity constraints," tech. rep., Dept. Computer Science, Ediburgh University, 1985.

[8] Bresson, X., Vandergheynst, P., and Thiran, J., "Multiscale active contours," *Int. J. Computer Vision*, vol. 70, no. 3, pp. 197–211, 2006.

[9] Brook, A., Bruckstein, A. M., and Kimmel, R., "On similarity-invariant fairness measures.," in *Scale-Space*, pp. 456–467, 2005.

[10] Bruckstein, A. M., Sapiro, G., and Shaked, D., "Evolutions of planar polygons," *Int. J. of Pattern Recognition and Artificial Intelligence*, vol. 9, pp. 991–1014, 1995.

[11] Bruckstein, A. M. and Netravali, A. N., "On minimal energy trajectories," *Comput. Vision Graph. Image Process.*, vol. 49, no. 3, pp. 283–296, 1990.

[12] Burger, M., "A framework for the construction of level set methods for shape optimization and reconstruction," *Interfaces and Free Boundaries*, no. 5, pp. 301–329, 2003.

[13] Burger, M. and Osher, S., "A survey on level set methods for inverse problems and optimal design," *European J. Appl. Math.*, vol. 16, 2005.

[14] Caselles, V., Catte, F., Coll, T., and Dibos, F., "A geometric model for active contours in image processing," Tech. Rep. 9210, Université Paris Dauphine, Ceremade, 1992.

[15] Caselles, V., Catte, F., Coll, T., and Dibos, F., "A geometric model for edge detection," *Num. Mathematik*, vol. 66, pp. 1–31, 1993.

[16] Caselles, V., Kimmel, R., and Sapiro, G., "Geodesic active contours," in *Proceedings of the IEEE Int. Conf. on Computer Vision*, (Cambridge, MA, USA), pp. 694–699, June 1995.

[17] Chan, T. and Vese, L., "Active contours without edges," *IEEE Transactions on Image Processing*, vol. 10, pp. 266–277, February 2001.

[18] Chang, H. and Valentino, D., "Medical image segmentation using a simulated charged fluid," in *Proceedings of SPIE*, vol. 5370, pp. 494–505, 2004.

[19] Charpiat, G., Keriven, R., Pons, J., and Faugeras, O., "Designing spatially coherent minimizing flows for variational problems baseed on active contours," in *ICCV*, 2005.

[20] Charpiat, G., Faugeras, O. D., and Keriven, R., "Approximations of shape metrics and application to shape warping and empirical shape statistics.," *Foundations of Computational Mathematics*, vol. 5, no. 1, pp. 1–58, 2005.

[21] Charpiat, G., Maurel, P., Keriven, R., Pons, J.-P., and Faugeras, O. D., "Generalized gradients: Priors on minimization flows," *Int. J. Computer Vision*, vol. 73, pp. 325–344, July 2007.

[22] Chen, Y., Tagare, H., Thiruvenkadam, S., Huang, F., Wilson, D., Gopinath, K., Briggs, R., and Geiser, E., "Using prior shapes in geometric active contours in a variational framework," *International Journal of Computer Vision*, vol. 50, pp. 315–328, Dec 2002.

[23] Chopp, D. L. and Sethian, J. A., "Motion by intrinsic laplacian of curvature," *Interfaces and Free Boundaries*, vol. 1, pp. 107–123, 1999.

[24] Cohen, L. D., "On active contour models and ballons," *Computer Vision, Graphics, and Image Processing: Image Processing*, vol. 53, no. 2, 1991.

[25] Crandall, M. and Lions, P. L., "Viscosity solutions of Hamilton-Jacobi equations," *Trans. Amer. Math. Soc.*, vol. 277, pp. 1–42, 1983.

[26] Cremers, D., Osher, S. J., and Soatto, S., "Kernel density estimation and intrinsic alignment for shape priors in level set segmentation," *International Journal of Computer Vision*, vol. 69, pp. 335–351, September 2006.

[27] Cremers, D. and Soatto, S., "A pseuso distance for shape priors in level set segmentation," in *IEEE Int. Workshop on Variational, Geometric and Level Set Methods*, pp. 169–176, 2003.

[28] Cremers, D., Tischhäuser, F., Weickert, J., and Schnörr, C., "Diffusion snakes: Introducing statistical shape knowledge into the mumford-shah functional.," *International Journal of Computer Vision*, vol. 50, no. 3, pp. 295–313, 2002.

[29] Delingette, H., "On smoothness measures of active contours and surfaces," in *VLSM '01: Proceedings of the IEEE Workshop on Variational and Level Set Methods (VLSM'01)*, (Washington, DC, USA), p. 43, IEEE Computer Society, 2001.

[30] do Carmo, M., *Riemannian Geometry*. Birkhäuser Boston, 1992.

[31] Droske, M. and Rumpf, M., "A level set formulation for the willmore flow," *Interfaces and Boundaries*, vol. 6, no. 3, pp. 361–378, 2004.

[32] Duci, A., Yezzi, A. J., Mitter, S. K., and Soatto, S., "Shape representation via harmonic embedding.," in *ICCV*, pp. 656–662, 2003.

[33] Freedman, M. H., He, Z. X., and Wang, Z., "Mobius energy of knots and unknots," *Annals of Mathematics*, vol. 139, no. 1, pp. 1–50, 1994.

[34] Fua, P. and Leclerc, Y. G., "Model driven edge detection," *Mach. Vision Appl.*, vol. 3, no. 1, pp. 45–56, 1990.

[35] Gage, M., "An isoperimetric inequality with application to curve shortening," *Duke Math. J.*, vol. 50, pp. 1225–1229, 1983.

[36] Gage, M., "Curve shortening makes convex curves circular," *Invent. Math.*, vol. 76, pp. 357–364, 1984.

[37] Gage, M. and Hamilton, R. S., "The heat equation shrinking convex plane curves," *Journal of Differential Geometry*, no. 23, pp. 69–96, 1986.

[38] Geman, S. and Geman, D., "Stochastic relaxation, Gibbs distribution, and Baysian restoration of images," *IEEE PAMI*, vol. 6, pp. 721–741, 1984.

[39] Goldenberg, R., Kimmel, R., Rivlin, E., and Rudzsky, M., "Fast geodesic active contours.," *IEEE Transactions on Image Processing*, vol. 10, no. 10, pp. 1467–1475, 2001.

[40] Grayson, M., "The heat equation shrinks embedded planes curves to round points," *J. Differential Geometry*, no. 26, pp. 285–314, 1987.

[41] Grayson, M., "Shortening embedded curves," *Ann. of Math*, vol. 129, pp. 71–111, 1989.

[42] Grenander, U., *General Pattern Theory*. Oxford University Press, 1993.

[43] Grenander, U. and Keenan, D. M., "On the shape of plane images," *SIAM J. Appl. Math*, vol. 53, pp. 1072–1094, 1991.

[44] Han, X., Xu, C., and Prince, J., "A topology preserving level set method for geometric deformable models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, pp. 755–768, June 2003.

[45] Han, X., Xu, C., and Prince, J., "Topology preserving deformable model using level sets," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, (Hawaii), pp. 765–770, December 2001.

[46] Hintermüller, M. and Ring, W., "An inexact newton-cg-type active contour approach for the minimization of the mumford-shah functional," *Journal of Mathematical Imaging and Vision*, vol. 20, no. 1, pp. 19–42, 2004.

[47] Horn, B. K. P., "The curve of least energy," *ACM Transactions on Mathematical Software*, vol. 9, pp. 441–460, Dec. 1983.

[48] Isard, M. and Blake, A., "Condensation – conditional density propagation for visual tracking," *IJCV*, vol. 1, no. 29, pp. 5–28, 1998.

[49] Jackson, J., Yezzi, A., and Soatto, S., "Tracking deformable moving objects under severe occulsions," in *IEEE Conference on Decision and Control*, Dec. 2004.

[50] Jalba, A., Wilkinson, M., and Roerdink, J., "Cpm: A deformable mode for shape recovery and segmentation based on charged particles," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 10, pp. 1320–1335, 2004.

[51] Kass, M., Witkin, A., and Terzopoulos, D., "Snakes: Active contour models," *International Journal of Computer Vision*, vol. 1, pp. 321–331, 1987.

[52] Kendall, D. G., "Shape manifolds, procrustean metrics and complex projective spaces," *Bull. London Math. Soc., 16*, vol. 16, 1984.

[53] Kichenassamy, S., Kumar, A., Olver, P., Tannenbaum, A., and Yezzi, A., "Gradient flows and geometric active contour models," in *Proceedings of the IEEE Int. Conf. on Computer Vision*, pp. 810–815, 1995.

[54] Kim, J., Fisher, J., Yezzi, A., Cetin, M., and Willsky, A., "Nonparametric methods for image processing using information theory and curve evolution," in *IEEE International Conference on Image Processing*, vol. 3, pp. 797–800, 2002.

[55] Kimia, B. B., Tannenbaum, A., and Zucker, S. W., "Shapes, shocks, deformations, i," *Int. J. Computer Vision*, vol. 15, pp. 189–224, 1995.

[56] Kolmogorov, V. and Boykov, Y., "What metrics can be approximated by geo-cuts, or global optimization of length/area and flux.," in *ICCV*, pp. 564–571, 2005.

[57] Lang, S., *Fundamentals of Differential Geometry*. Springer-Verlag, 1999.

[58] Le, H. and Kendall, D. G., "The riemannian structure of euclidean shape spaces: A novel envirnoment for statistics," *The Annals of Statistics*, vol. 21, no. 3, pp. 1225–1271, 1993.

[59] Leroy, B., Herlin, I., and Cohen, L., "Multi-resolution algorithms for active contour models," in *Proc. 12th Int. Conf. on Analysis and Optimization of Systems: Images, Wavelets and PDE's*, 1996.

[60] Leventon, M., Grimson, E., and Faugeras, O., "Statistical shape influence in geodesic active contours," in *IEEE Conf. on Comp. Vision and Patt. Recog.*, vol. 1, pp. 316–323, 2000.

[61] Lorenson, W. E. and Cline, H. E., "Marching cubes: A high resolution 3d surface construction algorithm," *ACM Computer Graphics*, vol. 21, no. 4, pp. 163–170, 1987.

[62] Ma, T. and Tagare, H., "Consistency and stability of active contours with euclidean and non-euclidean arc lengths.," *IEEE Transactions on Image Processing*, vol. 8, no. 11, pp. 1549–1559, 1999.

[63] Malladi, R., Sethian, J., and Vemuri, B., "Shape modeling with front propagation: a level set approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 17, pp. 158–175, 1995.

[64] Mansouri, A.-R., Mukherjee, D. P., and Acton, S. T., "Constraining active contour evolution via lie groups of transformation.," *IEEE Transactions on Image Processing*, vol. 13, no. 6, pp. 853–863, 2004.

[65] Mardia, K. V. and Dryden, I. L., "Shape distributions for landmark data," *Adv. Appl. Prob.*, vol. 21, no. 4, pp. 742–755, 1989.

[66] McInerney, T. and Terzopoulos, D., "Topologically adaptable snakes.," in *ICCV*, pp. 840–845, 1995.

[67] Mennucci, A. C. G., Yezzi, A., and Sundaramoorthi, G., "Properties of Sobolev–type metrics in the space of curves," *Preprint, arXiv:math.DG/0605017*, April 2006.

[68] Michor, P. and Mumford, D., "Riemannian geometries on the space of plane curves," *ESI Preprint 1425, arXiv:math.DG/0312384*, Dec. 2003.

[69] Michor, P. W. and Mumford, D., "An overview of the riemannian metrics on spaces of curves using the hamiltonian approach," *To appear in: Applied and Computational Harmonic Analysis (ACHA). arXiv:math.DG/0605009*, 2006.

[70] Mio, W. and Srivastava, A., "Elastic-string models for representation and analysis of planar shapes.," in *CVPR (2)*, pp. 10–15, 2004.

[71] Mumford, D. and Shah, J., "Boundary detection by minimizing functionals," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 1985.

[72] Mumford, D. and Shah, J., "Optimal approximations by piecewise smooth functions and associated variational problems," *Comm. Pure Appl. Math.*, vol. 42, pp. 577–685, 1989.

[73] Nain, D., Yezzi, A. J., and Turk, G., "Vessel segmentation using a shape driven flow.," in *MICCAI (1)*, pp. 51–59, 2004.

[74] Neuberger, J. W., *Sobolev Gradients and Differential Equations*. Lecture Notes in Mathematics #1670, Springer, 1997.

[75] Niethammer, M. and Tannenbaum, A., "Dynamic geodesic snakes for visual tracking," in *CVPR*, vol. 1, pp. 660–667, 2004.

[76] O'Hara, J., "Energy of a knot," *Topology*, vol. 30, no. 2, pp. 241–247, 1991.

[77] O'Hara, J., "Energy of knots," in *Ideal Knots* (Stasiak, A., Katritch, V., and Kauffman, L., eds.), pp. 288–314, 1998.

[78] Osher, S. and Sethian, J., "Fronts propagating with curvature-dependent speed: algorithms based on the Hamilton-Jacobi equations," *J. Comp. Physics*, vol. 79, pp. 12–49, 1988.

[79] Paragios, N. and Deriche, R., "Geodesic active contours and level sets for the detection and tracking of moving objects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 266–280, March 2000.

[80] Paragios, N. and Deriche, R., "Geodesic active regions: a new paradigm to deal with frame partition problems in computer vision," *International Journal of Visual Communication and Image Representation, Special Issue on Partial Differential Equations in Image Processing, Computer Vision and Computer Graphics*, vol. 13, pp. 249–268, June 2002.

[81] Paragios, N. and Deriche, R., "Geodesic active regions and level set methods for supervised texture segmentation," *International Journal of Computer Vision*, vol. 46, no. 3, p. 223, 2002.

[82] Peterfreund, N., "Robust tracking of position and velocity with Kalman snakes.," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 6, pp. 564–569, 1999.

[83] Polden, A., *Curves and Surfaces of Least Total Curvature and Fourth-Order Flows*. PhD thesis, Mathematisches Institut Unversitat Tubingen, Germany, September 1996.

[84] Rathi, Y., Vaswani, N., Tannenbaum, A., and Yezzi, A., "Particle filtering for geometric active contours and application to tracking deforming objects," in *IEEE CVPR*, 2005.

[85] Raviv, T. R., Kiryati, N., and Sochen, N., "Unlevel-set: Geometry and prior-based segmentation," in *Proc. European Conf. on Computer Vision*, 2004.

[86] Rochery, M., Jermyn, I., and Zerubia, J., "Higher order active contours and their application to the detection of line networks in satellite imagery," in *IEEE Workshop on VLSM*, October 2003.

[87] Rochery, M., Jermyn, I., and Zerubia, J., "Higher order active contours," Research Report 5656, INRIA, France, August 2005.

[88] Ronfard, R., "Region based strategies for active contour models.," *International Journal of Computer Vision*, vol. 13, pp. 229–251, October 1994.

[89] Rousson, M. and Paragios, N., "Shape priors for level set representations," in *Proc. European Conf. Computer Vision*, vol. 2, pp. 78–93, 2002.

[90] Rouy, E. and Tourin, A., "A viscosity solutions approach to shape-from-shading," *SIAM Journal on Numerical Analysis*, vol. 29, pp. 867–884, June 1992.

[91] Sapiro, G. and Tannenbaum, A., "Affine invariant scale space," *International Journal of Computer Vision*, vol. 9, no. 26, pp. 25–44, 1993.

[92] Sapiro, G. and Tannenbaum, A., "On affine plane curve evolution," *Journal of Functional Analysis*, vol. 1, no. 119, pp. 79–120, 1994.

[93] Sapiro, G. and Tannenbaum, A., "Area and length preserving geometric invariant scale-spaces.," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 1, pp. 67–72, 1995.

[94] Sethian, J. A., "A review of recent numerical algorithms for hypersurfaces moving with curvature dependent flows," *J. Differential Geometry*, vol. 31, pp. 131–161, 1989.

[95] Shah, J., "Shape recovery from noisy images by curve evolution," in *Int. Conf. on Signal and Image Processing*, 1995.

[96] Shah, J., "H^0-type Riemannian metrics on the space of planar curves," tech. rep., arXiv:math/0510192v1, 2005.

[97] Shi, Y. and Karl, W. C., "Differentiable minimin shape distance for incorporating topological priors in biomedical imaging," in *IEEE Int. Symp. Biomedical Imaging*, pp. 1247–1250, 2004.

[98] Siddiqi, K., Lauzière, Y. B., Tannenbaum, A., and Zucker, S., "Area and length minimizing flows for shape segmentation," *IEEE Transactions on Image Processing*, vol. 3, no. 7, pp. 433–443, 1998.

[99] Soatto, S. and Yezzi, A. J., "DEFORMOTION: Deforming motion, shape average and the joint registration and segmentation of images," in *ECCV (3)*, pp. 32–57, 2002.

[100] Sundaramoorthi, G. and Yezzi, A., "Global regularizing flows with topology preservation for active contours and polygons," *IEEE Transactions on Image Processing*, vol. 16, no. 3, pp. 803–812, 2007.

[101] Sundaramoorthi, G., Yezzi, A., and Mennucci, A., "Sobolev active contours," *Int. J. Computer Vision*, vol. 73, no. 3, pp. 345 – 366, 2007.

[102] Sundaramoorthi, G., Yezzi, A., and Mennucci, A., "Coarse-to-fine segmentation and tracking using sobolev active contours," *IEEE Transactions on PAMI*, To Appear.

[103] Sundaramoorthi, G., Yezzi, A., Mennucci, A., and Sapiro, G., "New possibilities with sobolev active contours," in *Scale Space and Variational Methods*, 2007.

[104] Terzopoulos, D. and Szeliski, R., *Active Vision*, ch. Tracking with Kalman Snakes. MIT Press, 1992.

[105] Trouve, A. and Younes, L., "Metamorphoses through lie group action," *Foundations of Computational Mathematics*, pp. 173–198, 2005.

[106] Tsai, A., Jr., A. J. Y., and Willsky, A. S., "Curve evolution implementation of the mumford-shah functional for image segmentation, denoising, interpolation, and magnification.," *IEEE Transactions on Image Processing*, vol. 10, no. 8, pp. 1169–1186, 2001.

[107] Tsai, A., Yezzi, A. J., III, W. M. W., Tempany, C., Tucker, D., Fan, A., Grimson, W. E. L., and Willsky, A. S., "Model-based curve evolution technique for image segmentation.," in *CVPR (1)*, pp. 463–468, 2001.

[108] Unal, G., Yezzi, A., and Krim, H., "Unsupervised texture segmentation by information-theoretic active polygons," *Int. J. Computer Vision*, vol. 62, no. 3, pp. 199–220, 2005.

[109] Unal, G. B., Krim, H., and Yezzi, A. J., "Fast incorporation of optical flow into active polygons.," *IEEE Transactions on Image Processing*, vol. 14, no. 6, pp. 745–759, 2005.

[110] Vese, L. A. and Chan, T. F., "A multiphase level set framework for image segmentation using the mumford and shah model," *Int. J. Computer Vision*, vol. 50, no. 3, pp. 271–293, 2002.

[111] Xu, C. and Prince, J. L., "Snakes, shapes, and gradient vector flow," *IEEE Transactions on Image Processing*, vol. 7, pp. 359–369, March 1998.

[112] Yezzi, A. and Mennucci, A., "Conformal riemannian metrics in the space of curves," in *MIA*, EUSIPCO04, 2004.

[113] Yezzi, A. and Mennucci, A., "Metrics in the space of curves," *Preprint, arXiv:math.DG /0412454*, May 2005.

[114] Yezzi, A., Tsai, A., and Willsky, A., "A statistical approach to snakes for bimodal and trimodal imagery," in *Int. Conf. on Computer Vision*, pp. 898–903, October 1999.

[115] Yezzi, A., Zllei, L. ., and Kapur, T., "A variational framework for joint segmentation and registration," in *IEEE CVPR - MMBIA*, 2001.

[116] Yezzi, A. J. and Mennucci, A., "Conformal metrics and true "gradient flows" for curves.," in *ICCV*, pp. 913–919, 2005.

[117] Younes, L., "Computable elastic distances between shapes," *SIAM J. Appl. Math.*, vol. 58, no. 2, pp. 565–586, 1998.

[118] Zhu, S. C., Lee, T. S., and Yuille, A. L., "Region competition: Unifying snakes, region growing, energy/bayes/MDL for multi-band image segmentation," in *ICCV*, pp. 416–, 1995.

VITA

Ganesh Sundaramoorthi was born in Edmonton, Alberta, Canada on July 4, 1982. In 2000 he began his undergraduate studies at Georgia Institute of Technology, Atlanta, USA, and in 2003, he obtained a B.S. in Applied Mathematics and a B.S. in Electrical Engineering. He continued his graduate studies at Georgia Institute of Technology. In 2005, he obtained a M.S. in Mathematics, and in Aug. 2007, he obtained the PhD in Electrical and Computer Engineering. His research interests are in the application of PDE and differential geometry to computer vision and image processing.